

A response to the UK Position Statement on forensic speaker comparison

Phil Rose & Geoffrey Stewart Morrison

School of Language Studies, Australian National University

philip.rose@anu.edu.au

geoff.morrison@anu.edu.au

1 Introduction

A recent issue of the *International Journal of Speech Language and the Law* contained a “position statement concerning the use of impressionistic likelihood terms in forensic speaker comparison cases” (French and Harrison 2007). This position statement was the result of a collaborative exercise among a number of researchers and forensic practitioners working in the United Kingdom. The foreword states that:

“the statement was circulated to all practising forensic speech scientists and interested academics within the UK. With one exception, all those contacted became co-signatories. The statement now reflects the all but unanimous position within the UK.” (p. 138)

The statement was also lodged with the prosecutorial bodies of Scotland, Northern Ireland, and England and Wales. For simplicity we will therefore refer to it as the *UK Position Statement*, with the proviso that it may not reflect the views of all interested parties in the UK¹, or have the force of law in any jurisdiction within the UK.

The editors of the *International Journal of Speech Language and the Law* invited responses to the UK Position Statement for publication in subsequent issues. This is one such response, a preliminary version of which was presented at the 17th meeting of the *International Association for Forensic Phonetics and Acoustics* in July 2008.

We first summarise the UK Position Statement as we understand it, and then present our response. It helps our exposition to precede our response with an outline of what we consider to be the correct framework for the presentation of forensic-voice-comparison evidence.

¹ Since we know of two “interested academics” in the UK who were not consulted, it is not the case that the UK Position Statement represents the views of all-but-one of the “practising forensic speech scientists and interested academics within the UK”.

2 Description of the UK Position Statement

It is made clear in the foreword that the UK Position Statement refers to comparison of voice recordings performed by experts, and thus relates to *technical* and not *naïve* forensic voice comparison (for these terms see Nolan 1983: 7, 1997: 744–745).

2.1 Motivations and goals

In its foreword, the UK Position Statement notes that it was motivated by a concern about “the framework in which conclusions are typically expressed in forensic speaker comparison cases” (p. 137). The awareness of there having been a problem with the existing framework is said to have initially arisen from the Appeal Court of England and Wales ruling in *R v. Doherty and Adams* ([1996] EWCA Crim 728) which involved the prosecutor’s fallacy related to the evidence-in-chief of a DNA expert.

The foreword to the UK Position Statement claims that it presents:

“... [a] new approach [which] brings about a fundamental change in the role of the analyst and the evidence. In the past forensic speech scientists were often thought of as identifying speakers. Within the new approach they do not make identifications. Rather, their role becomes that of providing an assessment of whether the voice in the questioned recording fits the description of the suspect.” (p. 138)

Footnote 2 of the UK Position Statement adds that the activity is therefore to be considered not identification, but comparison. The foreword also claims that the aim in developing the UK Position Statement was “... to bring the field [of forensic voice comparison] into line with modern thinking in other areas of forensic science” (p. 137), and that “This new framework is, at a conceptual level, identical to that used nowadays in the presentation of DNA evidence” (p. 138).

At the end of the document, the authors and signatories of UK Position Statement acknowledge that they :

“... accept in principle the desirability of considering the task of speaker comparison in a likelihood ratio (including Bayesian) conceptual framework. However, [they] consider the lack of demographic data along with the problems of defining relevant reference populations as grounds for precluding the quantitative application of this type of approach in the present context.” (p. 142, §6)

It might seem that the UK Position Statement’s rational eschewing of the likelihood-ratio-based framework in favour of its alternative proposal has a *faute de mieux* ring. However, this would be an unfair characterisation. It can be seen that attempts have clearly been made to manufacture a compromise with likelihood-ratio-based approaches, and it is thus possible to see this compromise – although it is not explicitly nominated as one – as a further motivation for the proposal. What we

will argue is that it is not the best compromise, and that it is not conceptually equivalent to the framework for the presentation of DNA evidence.

2.2 The UK Framework

A flow chart of the framework proposed in the UK Position Statement (henceforth *UK Framework*) is presented in Figure 1. In the UK Framework, speech samples are to be compared in terms of two serially ordered factors: *consistency* and *distinctiveness*.

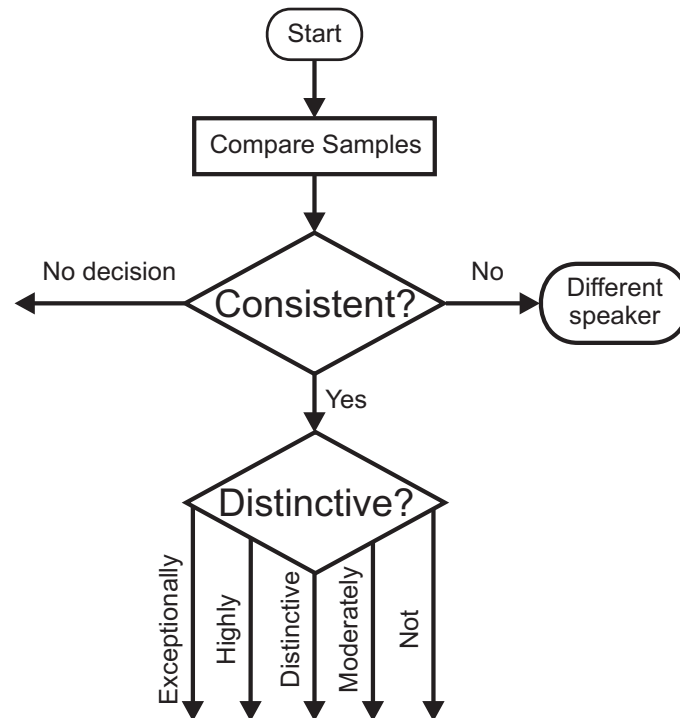


Figure 1. Flow chart representation of the UK Framework.

2.2.1 Consistency

Consistency is characterised as “whether the known and questioned samples are compatible, or consistent, with having been produced by the same speaker” (p. 141, §4.1). It is assessed by “the degree to which observable features [are] similar or different” (p. 141, §4.1). Differences between known and questioned samples count against consistency unless “they can be explained by models of acoustic, phonetic or linguistic variation (e.g. by reference to differential channel characteristics, [or within-speaker] sociolinguistic, psychological and/or physical factors)” (p. 141, §4.1). Consistency is quantified on a three-point scale: *consistent*, *not-consistent*, or *no-decision*. If *not-consistent* is returned, the samples are declared to have been spoken by different speakers. If *consistent* is returned, one proceeds to consider the question of distinctiveness (thus consistency and distinctiveness are serially ordered, and a judgment on distinctiveness is only made if there has first been a

positive determination on the issue of consistency). No instructions are provided as to actions following a *no-decision*.

2.2.2 *Distinctiveness*

The UK Position Statement emphasises that a positive determination of consistency does not imply that the known and questioned samples were necessarily spoken by the same person, since “the cluster of features leading to the consistency decision . . . [may] be shared by a substantial number of other people in the population” (p. 141, §4.2). It is implied that the likelihood that the samples have been produced by same speakers will be greater if their shared cluster of features is distinctive or unusual. Distinctiveness is assessed on a five-point scale ranging from *not-distinctive* to *exceptionally-distinctive*, the latter glossed with: “the possibility of this combination of features being shared by other speakers is considered to be remote” (p. 141, §4.2). There are no instructions as to how to proceed once a decision on distinctiveness has been reached. We presume that it is intended that the expert is to report that the samples are consistent with having been produced by the same speaker, and provide the determined degree of distinctiveness as an indicator of how unusual it would be to find this consistency if the two samples were not spoken by the same speaker.

3 Response to the UK Position Statement

First let us say that we applaud the motivation behind the UK Position Statement, and welcome its general direction. Specifically, we agree wholeheartedly with the goals of bringing the presentation of forensic-voice-comparison evidence into line with modern thinking in other areas of forensic science, and in particular of bringing it into line with modern practice in the evaluation of DNA evidence. These goals are not new: they were also proposed in Champod and Meuwly (2000), González-Rodríguez et al. (2006), González-Rodríguez et al. (2007), Rose (2002, 2003), Saks and Koehler (2005). We will argue, however, that the UK Position Statement fails to meet these goals.

Before our critical analysis of the UK Position Statement, it makes discussion easier if we first present what we consider to be the logically and legally correct framework for the evaluation of forensic comparison evidence: the likelihood-ratio framework. This is the framework which we believe represents the modern thinking in other areas of forensic science as exemplified by current practice in the evaluation of DNA evidence. Support for this position is provided by numerous textbooks, articles, and reviews written by forensic statisticians, legal experts, and forensic scientists, e.g. Aitken and Taroni (2004), Balding (2005), Champod and Meuwly (2000), Evett (1991, 1998), Friedman (1996), Good (1991), González-Rodríguez et al. (2006), Haigh (2005), Hodgson (2002), Lindley (1991), Robertson and Vignaux (1995).

3.1 The likelihood-ratio framework

The following account is abridged from Rose (2005: 49–54). It is written from the perspective that the objects of comparison are speech samples, but is in principle applicable to the evaluation of any type of forensic evidence (DNA, fingerprints, ballistics, toolmarks, etc.) where known and questioned samples are to be compared and it is possible to quantify physical properties which can vary across samples. For those interested in a fuller exposition of the likelihood-ratio framework, the following are recommended: Balding (2005), Lucy (2005), Robertson and Vignaux (1995), Rose (2002, 2003).

Typically in forensic voice comparison a recording of an unknown voice, usually of an offender, is to be compared with one or more recordings of a known voice, usually of a suspect or defendant. The interested parties (police/prosecution, trier-of-fact) want to know if the unknown (or questioned) voice comes from the same speaker as the known voice. They will usually understand that a definitive answer cannot be given; a trial is, after all, about making decisions in the face of uncertainty. So they will usually ask: *how probable is it that the samples have been said by the same person?* - a very reasonable way of putting it, since philosophers and statisticians will agree that the best way of quantifying uncertainty is by using probability (Lindley 1991). Implied, of course, will also be the rôle of evidence. That is, the question is really: *how probable is it, given the voice evidence, that the questioned and known samples have been said by the same person?* This is conventionally and conveniently formalised by the conditional probability expression at (1),

$$p(\mathbf{H}_{ss} \mid \mathbf{E}_{sp}) \quad (1)$$

where “ p ” stands for probability, “ \mathbf{H}_{ss} ” stands for the prosecution **H**ypothesis that the **s**ame **s**peaker was involved, the vertical stroke “ \mid ” stands for “given”, or “conditional upon”, and “ \mathbf{E}_{sp} ” stands for the speech evidence - the inevitably present differences between the suspect and offender speech samples. It is usual to use odds instead of probability, so the formal representation becomes (2),

$$p(\mathbf{H}_{ss} \mid \mathbf{E}_{sp}) / p(\mathbf{H}_{ds} \mid \mathbf{E}_{sp}) \quad (2)$$

where “ \mathbf{H}_{ds} ” stands for the **H**ypothesis that the samples were spoken by **d**ifferent **s**peakers.

The solution to this equation is given by Bayes’ Theorem, which has been known since at least the mid 1700s (Bayes, 1763). Bayes’ Theorem is of paramount importance when one wants to know the probability of a hypothesis given the evidence, and this is what gives it its special status in forensic identification. Bayes’ Theorem states informally that the probability of the hypothesis, given the evidence, can be estimated from two things: (1) how probable the hypothesis is, before the evidence is adduced; and (2) the strength of the evidence.

The odds form of Bayes’ Theorem, applied to forensic voice comparison, is given at (3). It says that the odds in favour of it being the same speaker, given the speech evidence (this is what

everyone wants to know and is called the posterior odds and is at the left of the equals sign), is the prior odds in favour of it being the same speaker times the strength of that evidence. Thus the odds in favour of it being the same speaker can be calculated from two terms: the *prior odds* and the *likelihood ratio*.

$$\frac{p(H_{ss} | E_{sp})}{p(H_{ds} | E_{sp})} = \frac{p(H_{ss})}{p(H_{ds})} \times \frac{p(E_{sp} | H_{ss})}{p(E_{sp} | H_{ds})} \quad (3)$$

Posterior Odds *Prior Odds* *Likelihood Ratio*

The *prior odds* are the odds in favour of the hypothesis before the voice evidence is adduced. These are simply the probability that it is the same speaker divided by the probability that it is a different speaker. In its limit, it could be anyone in the world, but the prior odds can usually be considerably narrowed-down by taking into account obvious information in the voice like sex and accent, as well as other pragmatic information.

The *likelihood ratio* is the most important metric in forensic voice comparison because it is a measure of the *strength of the evidence* in favour of a hypothesis, and it is what the expert should try to estimate. The formula at (3) shows that the likelihood ratio too is a ratio of probabilities, but these probabilities are probabilities of *evidence*, not *hypotheses*. The likelihood ratio quantifies how much more likely you are to get the differences between the suspect and offender speech samples assuming they have come from the same speaker than assuming they have come from different speakers.

If you are more likely to get the speech evidence assuming that the samples came from the same speaker than from different speakers - if $p(E_{sp} | H_{ss})$ is greater than $p(E_{sp} | H_{ds})$ - that counts as support for the prosecution claim that the samples came from the same speaker. If, on the other hand, you are more likely to get the speech evidence assuming that the samples came from different speakers than from the same speaker - if $p(E_{sp} | H_{ds})$ is greater than $p(E_{sp} | H_{ss})$ - that counts as support for the defence claim. If you are just as likely to get the evidence assuming same-speaker as different-speaker provenance - if the ratio of $p(E_{sp} | H_{ss})$ to $p(E_{sp} | H_{ds})$ is one - the evidence is useless.

Thus the magnitude of the likelihood ratio quantifies the strength of the evidence: greater than unity means support for same-speaker claim; less than unity means support for different-speaker claim; unity (or values close to it) mean evidence is useless (or next to useless).

The main textbooks on the evaluation of forensic evidence (e.g. Robertson and Vignaux, 1995) and forensic statistics (e.g. Aitken and Stoney 1991, Aitken and Taroni 2004, Lucy 2005), stress that it is the role of the forensic expert to quantify the strength of the evidence by estimating

its likelihood ratio: the probabilities of the evidence under competing prosecution and defence hypotheses:

“The case made for this approach, whether the subject matter is DNA, glass fragments, clothing fibres or whatever, is overwhelming” (Haigh 2005)

“Statistical evaluation, and particularly Bayesian methods such as the calculation of likelihood ratios . . . are the only demonstrably rational means of quantifying the value of evidence available at the moment: anything else is just intuition and guess-work.” (Lucy 2005: 138)

3.2 Comparison, not identification

The UK Position Statement emphasises that its proposal is not to be considered identification, but comparison, and this is the term they use (and we have adopted). This choice is worth commenting on, because many terms are currently in use: *identification*, *recognition*, *verification*, and *discrimination*. In parts of the literature some of these terms, e.g. *identification* and *recognition*, have been used interchangeably, and in other parts different terms, e.g. *identification* and *verification*, have been used to designate logically different types of analysis or different applications (Rose 2002: Chapter 3). We agree with the UK Position Statement that *comparison* is the most appropriate term, but for slightly different reasons. Terms like *identification*, *verification*, and *recognition* imply the expression of a posterior probability (i.e., the probability that the suspect and offender voices are the same), and some have connotations of providing a categorical decision. As we argue below, and was stated in Rose (2002: 89), since it is logically not possible, and legally inappropriate, for a forensic expert to provide a posterior probability, there is no identification, verification or recognition involved, and we are therefore in agreement with the UK Position Statement that with respect to forensic work these terms should be eschewed in favour of a neutral term such as *comparison*, which does not carry connotations of providing a posterior probability decision. We would also suggest that since the objects of comparison are recordings of voices – voices are compared, speakers are not – we adopt the term *forensic voice comparison* rather than the UK Position Statement’s *forensic speaker comparison*. Nolan (1983, 1996) gives the best current characterisation of a voice for forensic purposes. His semiotic account is also described in detail in Rose (2002: Ch. 10).

3.3 Prohibition on probability of hypothesis, given evidence

The first important proposal of the UK Position Statement is the recommendation that the expert refrain from giving the probability of hypothesis, given evidence [$p(H|E)$]. We strongly endorse this: for some time now this has indeed been the position adopted by forensic statisticians and more recently by some courts (see the historical discussion in Aitken and Taroni 2004: 108, 122–128, 153–155, 208–213, Balding 2005: 145–153). However, the UK Position Statement implies that the

reason a forensic expert should not quote $p(H|E)$ is because this gives a “false weighting” to the evidence which it relates to the prosecutor’s fallacy. The prosecutor’s fallacy refers to the erroneous transposing of evidence with hypothesis, i.e. replacing $p(E|H)$ with $p(H|E)$. This is the same as saying that because the evidence is 1000 times more likely under an assumption of guilt, the defendant is 1000 times more likely to be guilty (Aitken and Taroni 2004: 79–82, Balding 2005: 146–147, Donnelly 2005, Evett 1998, Thompson and Schumann 1987). Certainly forensic scientists should avoid making this error themselves, and should do everything within their power to prevent their testimony from being misinterpreted by lawyers, judges, and juries. But the UK Position Statement’s argument about “false weighting” does not address the real reasons why forensic scientists must provide the probability of evidence given hypothesis [$p(E|H)$], and why they cannot provide the probability of hypothesis given evidence [$p(H|E)$].

There are two reasons why the forensic expert cannot provide the probability of the hypothesis given the evidence: one logical, one legal. The logical reason follows trivially from Bayes’ Theorem. The posterior odds are determined by two things: the strength of the evidence (likelihood ratio) and the prior odds (see equation 3). The expert is not privy to the prior odds, hence a posterior cannot logically be quoted. The legal reason has to do with violations of the ultimate issue rule: In cases where the offender sample is truly incriminating, the expert’s pronouncement that the suspect is likely to have said the incriminating speech is equivalent to an expression of probable guilt, and this usurps the role of the trier-of-fact (the judge or the jury, depending on the legal system).

Although the UK Position Statement appears to condemn the practice of providing probability of hypothesis given evidence, there are two places where providing $p(H|E)$ is in fact recommended. These are discussed within the following two subsections, one related to differences between DNA and speech data, the other with closed-set comparisons.

3.3.1 Differences between DNA and speech data

The first violation of the prohibition on quoting $p(H|E)$ occurs in relation to the *not-consistent* determination on the issue of consistency:

“Where the samples are not consistent we see no logical flaw in making the statement that the samples are spoken by different speakers. This may be stated with a degree of confidence appropriate to the exigencies of the data.” (p. 141, §4.3)

Saying, with a given confidence, that the samples were spoken by different speakers because they are not consistent is a $p(H|E)$ statement. *Pace* the UK Position Statement claim above, by Bayes’ Theorem it is, in fact, a logical flaw.

We suspect that this inconsistency has crept in because the authors of the UK Position Statement were attempting to adapt a model for DNA analysis without taking into account impor-

tant differences in the nature of DNA versus speech evidence. Although the evaluation of forensic speech evidence can indeed be done in the same way as DNA, with likelihood ratios – this was demonstrated in a recent paper with both automatic and traditional approaches (González-Rodríguez et al. 2007) – one must be careful when drawing comparisons between DNA and forensic voice data. This is because of differences in the nature of the variation involved. The three aspects of variation which are of the greatest importance in forensics are the **type** of variation involved; how many **levels** of variation to account for; and the **magnitude** of the variation. DNA differs from speech in all three, but here it is the type and levels of variation that are of importance.

Variables can be either continuous or discrete, or a combination. DNA variables – typically the length of STR alleles at given loci - are discrete. With discrete variables it is possible to talk about a match, for example that both samples show a genotype having the same combination of 14, 16 at the D18 locus and 9.3, 9.3 at THO1 (Balding 2005: 3) [9.3, although it looks continuous, is not. It means that one of the ‘repeats’ only has three out of the expected four bases.] A non-match is also possible with DNA.

Whilst DNA evidence is discrete, speech evidence is continuous: cepstral coefficients, formant centre-frequencies, etc. are continuously valued variables, and even higher-level features such as the incidence of a particular allophone result in continuously valued proportions. Also, whereas the properties of speech vary from occasion to occasion, the DNA of a biological organism will be the same every time it is measured (making allowance for measurement error, contamination, somatic changes, transplants, chimera etc.).

Allowing for caveats, then, the properties of categoricity and invariance mean that if two DNA profiles do not match, the probability of getting this assuming that they have come from the same organism is zero. In this case the numerator of the likelihood ratio is zero and the posterior probability that they have come from the same organism, irrespective of the priors, is also zero. Allowing for caveats, DNA can therefore be used to provide definitive evidence of exclusion. Not so speech. In general, speech data do not, by their nature, allow such a definitive exclusion. We can imagine some conditions under which a voice comparison could result in a definitive exclusion, e.g., the vocal tract of a young child could not produce the lower formants of a typical adult male, but in such cases the voices are likely to sound so different that it would be highly unlikely that a forensic expert would be consulted.

Again allowing for caveats, given a match between two DNA profiles, the probability of observing this assuming that both samples come from the same organism is one (Aitken and Taroni 2004: 404, Evett 1998). With the numerator of the likelihood ratio unity, its magnitude is dependent on the size of its denominator. The denominator is the *random-match probability* (referred to in the

UK Position Statement as the *random occurrence ratio*²). This is the probability of getting a match with the obtained DNA profile if one randomly samples profiles from members of the relevant population. Since under such circumstances the likelihood ratio is equivalent to the inverse of the random-match probability, strength of DNA evidence can also be presented in the form of the random-match probability rather than the likelihood ratio.

It may be the case that an inappropriate transference from DNA analysis in the UK Position Statement has also resulted in problems involving the concept of random match. The UK Position Statement considers a case where, in the UK, a DNA match between the suspect and offender is established, with a random match probability of 1 in a million (i.e. one person in a million has a profile matching that of the offender), and that there are 60 million people in the UK. Under these circumstances the UK Position Statement quotes "... a one in sixty chance that the DNA came from the defendant" (p. 139).³ The UK Position Statement continues:

"The estimation that 1 person in a million will share the DNA profile is known as its 'random occurrence ratio'. Phoneticians can calculate the random occurrence ratio for very few features of speech. Exceptions are fundamental frequency (a measure of voice pitch), articulation rate (speed of speaking) and stammering." (p. 140, §3)

Since speech data are inherently continuous and it is a truism that a speaker never says exactly the same thing the same way twice, there is always variation between speech samples, and the numerator of a likelihood ratio derived from a forensic voice comparison can never be zero or one. The concept of random match is thus not applicable to continuously valued speech data. The strength-of-evidence from a forensic speaker comparison on continuously valued data can only be expressed in the form of a likelihood ratio.

Whereas random-match probability is certainly a meaningless concept with respect to inherently continuously-valued properties such as fundamental frequency, arguably it may be possible to calculate random-match probabilities for incidence of speech features such as stammering. However, this would only be under the assumptions that a recording of someone who habitually stam-

² "The term 'random occurrence ratio' introduced by the court [in *R v. Dohney and Adams*] appears to be a synonym for match probability. This novel coinage is an unwelcome addition to the many terms already available: its unfamiliarity could confuse." Balding (2005: 152)

³ The correct answer is actually a one in sixty-*one* probability that the defendant is the source of the DNA trace (or *odds* of 60 to one against). Of the 60 million possible perpetrators in the UK, one is guilty and the remaining 59,999,999 are innocent. The guilty party will provide one match, and the 59,999,999 others will provide 60 matches (because the probability of a random match is 1/1,000,000, and $59,999,999 \times (1/1,000,000) = 60$ (to the nearest integer)). So there will be in total 61 possible matches. Out of this 61, one is the true match, and the rest are false positives, so the probability that the suspect left the trace is 1/61. A simplified formula for calculating the probability of guilt under these circumstances [$P(G|E) = 1/1+N*p$] is given by Balding (2005: 11), where $P(G|E)$ stands for the probability of **G**uilt, given the **E**vidence, N is the number of people *other than the suspect* that could have been the perpetrator, and p is the probability of a random match.

mers will always contain instances of stammering, and a recording of someone who does not generally stammer will never contain instances of stammers.

3.3.2 Closed-set comparisons

The second violation of the prohibition on giving $p(H|E)$ occurs in section 5 of the UK Position Statement:

“In a minority of cases, however, there is independent evidence (e.g. video surveillance) to show that a closed set of known speakers was present and participating in the conversation. In such cases the comparison task becomes an issue of who said what. In these circumstances, if the voices are sufficiently distinct from one another, we consider it justified to make categorical statements of identification.” (p. 142, §5)

Making a categorical statement of identification, given sufficient distinctness, is a *probability of hypothesis, given evidence* statement, and a logical violation of Bayes’ Theorem. Closed set comparisons can be treated in exactly the same way as open, from the point of view of the strength of evidence (see Rose 2002: 64, 74).

3.4 Two-stage assessment

Another important part of the proposal, and again a welcome step in the right direction, is its bipartite assessment in terms of consistency and distinctiveness. Section 4.2 of the UK Position Statement is correct in assuming that the value of the evidence is dependent not only on the *similarity* between the two samples, but also on how *typical* they are. Two very similar, yet typical, samples will not be valued as highly in terms of strength of evidence in favour of identity as two very similar, yet atypical, samples. This is a point not always understood, and it is not uncommon to encounter the assumption that identity follows from similarity alone. It is therefore good to see this made clear in the UK Position Statement.

At first glance the UK Framework’s *consistency* and *distinctiveness* terms appear to parallel the *numerator* and *denominator* of the likelihood ratio discussed in section 3.1. However, the UK Framework’s use of a bipartite assessment via *consistency* and *distinctiveness* is not equivalent to the calculation of a likelihood ratio. An essential feature of a likelihood ratio is that the numerator and denominator are measured on the same scale (they are both values from probability densities) and are directly associated with each other (i.e. in the form of a ratio). In the UK Framework, consistency and distinctiveness are serially ordered; are measured on different scales (one has three discrete levels and the other five); and they are not directly related to each other.

The trier-of-fact needs to know whether the differences between the speech samples are more likely to have arisen if they were spoken by the same speaker, or more likely to have arisen if they were spoken by different speakers, or equally likely to have arisen irrespective of whether they

were spoken by the same speaker or by different speakers. It is not possible to do this unless both terms are quantified on the same scale and are directly related to each other. This we therefore consider a weakness of the bipartite UK Framework.

The UK Framework's two-stage analysis of *consistency* and *distinctiveness* is in fact reminiscent of Evett's (1977) evaluation of evidence in terms of *comparison* and *significance* stages. In this approach one first decides if there is a match on the basis of prior agreed criteria – say both samples lie within three standard deviations of each other. Then one assesses the probability of finding the observed degree of similarity in the relevant population (see criticism of this framework in Aitken and Taroni 2004: 10–11, and in Evett 1991). Although several variants of two-stage assessment have historically been applied to DNA evidence, this was superseded by likelihood-ratio evaluation, and two-stage approaches are not representative of modern practice (see Foreman et al. 2003, for a historical review of the interpretation of DNA evidence up to that point in time).

In addition to the systematic problems with the two-stage assessment of the UK Framework, there are problems with the structure of its component parts. Below we discuss the cliff-edge problem and multivariate data problem, both related to the fact that *consistency* and *distinctiveness* in the UK Framework have a finite number of categorical outcomes. We discuss these problems in relation to *consistency*, but it should be clear that near identical arguments can be made with respect to *distinctiveness*. The division into three versus five categorical outcomes is immaterial to the logic of the arguments. We also discuss the unfortunate choice of the word “consistent”.

3.4.1 *The cliff-edge effect*

It is a phonetic truism that there are always differences between speech samples, even if they come from the same speaker repeating the same thing only seconds apart on the same occasion. Furthermore, these differences will be gradient, not categorical, and as such the difference between the samples cannot be adequately captured by the UK Framework's ternary categorical outcome.

Consider vowel formants as typical gradient features. The situation often arises where a set of vowels from the suspect and offender are compared with respect to their formant centre-frequency distributions. Table 1 gives mean and standard-deviation values calculated from the formants of tokens of a single vowel phoneme taken from two intercepted telephone conversations. For the sake of argument let us assume that it has been determined that that the distributions of the features are sufficiently close to normal to warrant using the mean and standard deviation as appropriate representations of central tendency and dispersion.

Table 1. Means and standard deviations for formant centre frequencies measured in hertz for 15 tokens of Australian English /ə:/ in two different intercepted telephone conversations.

	F2	F3
Suspect		
mean	1429	2298
standard deviation	30	67
Offender		
mean	1450	2329
standard deviation	48	56
Difference between means	21	31

Consider F2 as a single feature. We imagine that, even in the absence of specific population data, on the basis of their experience most practitioners would agree that 21 Hz (a 1.5% difference) is well within the size of difference that might be expected for schwa F2 means from the same speaker talking normally on different occasions. If working in the UK Framework, they would return a decision of *consistent*. Since the UK Framework has not specified how one would arrive at a decision on consistency, we have to assume it would be by implicit reference to some quantification of variation such as the feature’s standard deviation. Table 1 shows that for both the suspect and offender samples the standard deviation in F2 is larger than the difference in F2 means between the samples. But at what point does one change from deciding that the observed values of the feature are consistent to deciding that they are inconsistent with having come from the same speaker? Should the boundary be at two standard deviations, or perhaps three? Should one apply a frequentist statistical test, such as a *t*-test with a prescribed alpha-level of 0.05, or perhaps 0.01? Such an approach imposes a categoricity, with an attendant cliff-edge effect (Robertson and Vignaux 1995: 118). If the boundary were set at two standard deviations using the standard deviation from the suspect data, a difference in the F2 means of 59.9 Hz would be ruled *consistent* but a difference of 60.1 Hz would be ruled *not-consistent* (or would both be declared *no-decision*). Should the decision hang on a difference of 0.2 Hz? We contend that any metric of similarity for use with forensic speaker comparison should be a gradient not categorical.

3.4.2 Problems with multivariate data

A further problem with the UK Framework’s consistency factor arises from the necessity of comparing samples with respect to multiple features. The foreword to the UK Position Statement describes the process of forensic voice comparison as:

“[involving] ‘separating out’ the samples into their constituent phonetic and acoustic ‘strands’ (e.g., voice quality, intonation, rhythm, tempo, articulation rate, consonant and vowel realisations) and analysing each one separately.” (p. 138)

No problem there, since multidimensionality is one of the things that contributes to the forensic discriminability of voices. However, the Statement gives no indication of how ultimately to combine the individual evidence from each of these ‘strands’. Returning to the formant data in Table 1 and the example of a two-standard-deviation boundary based on the standard deviation from the suspect data (the argument would apply equally well, irrespective of the features under consideration or of the prescribed threshold). What would be the decision with respect to consistency if the difference in the F2 means were 50 Hz, within the boundary, but the difference in the F3 means were 140 Hz, outside the boundary? What if a pair of samples were judged *consistent* on nine features but *not-consistent* on one? Again it is incumbent upon the UK Framework to specify how one would proceed under such circumstances. Here, as elsewhere, an example of the approach would have been useful, as it appears to us to be very difficult to implement. If one eschews categorical quantification, then one can simply use a multivariate gradient metric of similarity, such as the numerator of the multivariate likelihood ratio. The literature already contains many procedures for dealing with multivariate data within the likelihood-ratio framework (e.g. Aitken and Lucy 2004, Pigeon, Druyts, and Verlinde 2000, Reynolds, Quatieri, and Dunn 2000).

3.4.3 *The semantics of ‘consistent’*

Section 4.1 of the UK Framework defines consistency as “the degree to which observable features [are] similar or different” (p. 141, §4.1). This is certainly a coherent notion in that it is possible to estimate how similar the questioned voice is to the known voice in the specified feature(s). However, as it stands, consistency is epistemologically very weak and its implementation is problematic and far from clear. In particular, the choice of the word *consistency* to represent this parameter is not felicitous. In their book on the evaluation of evidence, Robertson and Vignaux strongly criticize its use by forensic experts:

“Worst of all is the word ‘consistent’, a word in unfortunately common use by forensic scientists, pathologists and lawyers. ... Unfortunately for clear communication ... lawyers usually interpret ‘consistent with’ as meaning ‘reasonably strongly supporting’.”

Robertson and Vignaux (1995: 56)

We suspect that juries, and perhaps even judges, are also likely to construe *consistent with coming from the same speaker* as meaning *likely to have come from the same speaker*, although we think that within the UK Framework it is clear that this is not what is intended by the term *consistent*. Robertson and Vignaux also point out, in its proper (though not usual) sense, *consistent* has little epistemological force, as “consistent with H” implies nothing about the likelihood that H is true (*the facts are consistent with H but unlikely to result from H* is coherent in this sense of *consistent*).

3.5 Problems with populations and samples

As stated above, the UK Position Statement says that a quantitative likelihood-ratio-based evaluation of evidence of the type outlined in section 3.1, however desirable, is not possible due to two factors: “problems of defining relevant reference populations”, and “lack of demographic data” (p.142 §6). We readily acknowledge that these are real problems: probably the most pressing at the moment. The first is theoretical and relates to the choice of the relevant population to sample, and the size of that sample; the second is practical and relates to the actual collection of data. Below we address both in turn. We would argue that these problems, however real, do not prevent the use of likelihood ratios.

3.5.1 *The appropriate reference population*

One of the problems for likelihood ratio-based approaches mentioned by the UK Position Statement concerns “defining relevant reference populations.” In order to estimate the strength of evidence with a likelihood ratio a *reference*, or *background* sample from the relevant population is needed.⁴ The UK Position Statement is certainly correct in assuming this is a problem, and not only for speech: it continues to be a challenge in the evaluation of DNA samples (Aitken 1991). The choice of the appropriate population to sample is strictly speaking dependent on the alternative hypothesis. In a typical case the prosecution will contend that the voice in the questioned sample is the same as that in the known sample, i.e., both are the voice of the defendant. The defence will contend that the voice in the questioned sample is not that of the defendant. However, the contention is unlikely to be simply that the speaker of the questioned sample is some other human being. Rather the claim is more likely to be implicitly that the voice is that of another speaker of the same sex as the defendant who speaks the same dialect of the same language. The defence hypothesis could be even more restricted: that the questioned voice is that of someone who (to a naïve listener such as a police officer or a lawyer) sounds like the defendant; or it could be that the source of the questioned voice is the defendant’s brother; or even identical twin, should they have one (see the discussion in Lucy 2005: 129–133, with respect to population limits). It is likely that, at the time of conducting the analysis, the forensic-voice-comparison expert does not know the specifics of the defence hypothesis, and must therefore anticipate what it might be. Decisions on the relevant reference population will have to be made on a case-by-case basis (Evetts and Weir 1998: 44–45; see also discussion in Aitken 1991).

A commonly asked question about reference population sampling, and one which may have been implied in the UK Position Statement’s observations about populations, is: how big should the

⁴ Unfortunately in the literature the terms *reference population* or *background population* have sometimes been used when what is actually meant is a *sample of the population*. A reference or background population [sic] is not the same as the population of all possible perpetrators (the latter may be adduced in the estimation of the prior odds).

sample be? This is a very important question, both from the point of view of research and in court (since it is one of the things that needs to be justified). The answer is that it depends on the precision required. Aitken (1991) discusses some approaches to solving this problem.

3.5.2 *Lack of information on distribution*

By *lack of demographic data* we assume, on the basis of the following quote, that what is meant is a lack of knowledge of the distribution of typical forensic comparison features in the population:

“for the overwhelming majority of voice and speech features examined in casework, it is simply not known how widely they are distributed in the population.” (p. 140, §3)

The UK Position Statement echoes Rose (2002: 78) in pointing out the consequences of this - that if one doesn't know the incidence of a feature in the population it is not possible to quote the probability of observing it at random from that population, thus: “forensic phoneticians are unable to provide numerical statements of probability” (p. 140, §3).

Clearly it is true that if one does not have access to estimates of the distribution of speech properties in the relevant population then one cannot quantitatively estimate a likelihood ratio (with appropriate confidence limits/credible intervals) for a forensic voice comparison. However, by the same token it must also follow that without such estimates one also cannot make decisions with respect to the UK Framework's *distinctiveness* factor. It is not made clear how this inconsistency is to be resolved.

Perhaps the authors and signatories of the UK Position Statement have envisioned that *distinctiveness* judgments could be made by an expert based on their own experience, but then such an inference would also enable a likelihood ratio estimate: a feature value judged in the expert's opinion as “exceptionally distinctive” can also be characterised as having a low probability in the population (the denominator of the likelihood ratio).

Section 6 of the UK Position Statement concludes with:

“However, we consider the lack of demographic data along with the problems of defining relevant reference populations as grounds for precluding the quantitative application of this type of approach in the present context. In view of these difficulties, the framework we endorse is the one set out in 4 above.” (p. 142, §6)

We must assume that “present context” should be read as “present context in the UK”, although this is not made entirely clear. The immensely complex current linguistic situation in the UK is such that, given the identification of an accent in the offender and suspect samples, there is at present no, or next to no, information available on the distribution of (acoustic) features in that accent. Under these circumstances, neither quantitative likelihood ratio nor *distinctiveness* estimates are possible, unless the forensic expert goes and collects samples from the population. However, in Australia and

Spain forensic scientists have collected distribution data and presentation of forensic-voice-comparison evidence in the form of likelihood ratios has been tendered as reports and received by the courts. We acknowledge that this is facilitated by the fact that variation in Australian English accents appears to be much less than in English accents within the UK.

In the UK context the problem remains that if appropriate databases of speech recordings are not available, then the forensic expert will have to collect and analyse them. With a fully automatic system, the analysis is a relatively cheap operation, but for traditional acoustic-phonetic approaches a major investment in human labour is required to collect and analyse reference data. We note that further work is already underway in the UK on the collection of databases and measurement of features therein (Nolan et al. 2006, Hudson et al. 2007, Clermont et al. 2008); however, it looks as if the short-term reality in the UK will be that quantitative estimates of strength of evidence will not be possible in some, perhaps many, cases.

If a short-term solution is to present qualitative estimates of strength of evidence, as opposed to quantitative ones, then it would be better to make such statements in the form of a likelihood ratio, rather than using the UK Framework. Such an approach is recommended by Robertson and Vignaux, and Jessen, if population distribution data are not available:

“To assess a likelihood ratio it is not essential to have precise numbers for each of the probabilities. The value of the evidence depends upon the ratio of these numbers. Therefore, if we believe that the evidence is 10 times more probable under one hypothesis than the other, the likelihood ratio is 10, whatever the precise values of the numerator and denominator may be. Often we will be able to assess this ratio roughly on the basis of our general knowledge and experience.” Robertson and Vignaux (1995: 21)

“Even in areas where no such population statistics exist, and therefore no quantification is possible, the Bayesian approach should be used as a conceptual framework that provides the logical backbone of voice comparison analysis.” (Jessen 2008:13)

Clearly, many of the UK practitioners are highly competent phoneticians with extensive experience in comparing voices, forensically and otherwise, with respect to many features. They can be expected to have a good intuition about the expected magnitudes of both within-speaker and between-speaker differences in these features. Thus they could be expected to be able to proffer an expert opinion of the following kind: “From my experience I think you would be much more likely to get the differences I have listed between the offender and suspect speech samples assuming that they had come from the same speaker, rather than different speakers.” (or *mutatis mutandis*). A qualitative statement like this, of the probability of evidence under competing hypotheses, surely would have some value to the court, and would also be consistent with our arguments for the likelihood ratio framework as the logically correct framework for the estimation of strength of evidence.

We are however in agreement with Lucy that a qualitative statement remains a poor substitute for a quantitative likelihood ratio (and its credible interval/confidence limits):

“It would not be ideal, nor desirable, to use such an estimate to evaluate a key piece of evidence in a major criminal trial” (Lucy 2005: 137).

4 Summary & Conclusion

This response has critiqued the proposals outlined in the UK Position Statement for a change in approach to forensic speaker comparison. The authors and signatories of the UK Position Statement deserve credit for acknowledging the need for a change, and initiating one, in the UK. Specifically, we see as positive the Statement’s recommendations to avoid the logically problematic traditional conclusions couched in terms of probability of hypothesis, given evidence; and to take into account both similarity and typicality of the speech samples under comparison. We also approve of the use of the term *comparison*, instead of *identification*, *verification*, or *recognition*.

We have identified three basic weaknesses in the proposal. Firstly, the apparent treatment of speech as if its features were discrete and invariant, like DNA, has the result that the approach would be very difficult to implement. Secondly, we pointed out the inconsistencies in first prohibiting but then allowing *probability of hypothesis, given evidence* statements; and proposing distinctiveness statements while acknowledging the absence of information upon which to base them. Thirdly, the absence of a way of relating the consistency and distinctiveness assessments does not help the trier-of-fact to interpret them.

Given the UK Position Statement’s in-principle endorsement of the likelihood ratio approach, we have tried to criticise it on its own merits, and to avoid criticising it for not being a likelihood ratio framework. However, the bipartite assessment is not a likelihood ratio, and it is the use of likelihood ratios that characterises modern thinking on the evaluation of forensic comparison evidence. We would therefore reject the claim in the Statement’s foreword that the UK Framework is “at a conceptual level, identical to that used nowadays in the presentation of DNA evidence” (p. 138). Consequently, we would also argue that the UK Position Statement has not achieved its goal, however laudable, of “... bring[ing] the field [of forensic voice comparison] into line with modern thinking in other areas of forensic science” (p. 137).

It will be interesting to see the proposed UK Framework implemented in research and case-work. We, of course, would encourage forensic-voice-comparison researchers and practitioners world-wide to rapidly move towards adopting quantitative likelihood-ratio statements as standard (although there will be some features of forensic speech samples for which it will only ever be possible to give qualitative estimates). Given the amount of energy in the UK now going into acquiring

quantitative data to underpin likelihood ratio-based forensic voice comparison, we hope this may actually soon be the case in the UK.

References

- Aitken, C. G. G. and Stoney, D. A. (1991) *The Use of Statistics in Forensic Science*. Chichester, UK: Ellis Horwood.
- Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Forensic Evidence for Forensic Scientist* (2nd ed.). Chichester, UK: Wiley.
- Aitken, C.G.G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53(4):109-122.
- Balding, D. J. (2005) *Weight-of-evidence for Forensic DNA Profiles*. Chichester, UK: Wiley.
- Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418.
- Champod, C. and Meuwly, D. (2000) The inference of identity in forensic speaker recognition. *Speech Communication*, 31: 193–203.
- Clermont, F., French, J. P., Harrison, P., and Simpson, S. (2008) Population data for English Spoken in England. Paper presented at the 17th meeting of the International Association for Forensic Phonetics and Acoustics, Lausanne, Switzerland, July 2008.
- Donnelly, P. (2005) Appealing statistics. *Significance*, 2(1): 46–48.
- Evetts, I. W. (1977) The interpretation of refractive index measures. *Forensic Science International*, 9: 209–217.
- Evetts, I. W. (1991) Interpretation: A personal odyssey. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 9–22. Chichester, UK: Ellis Horwood.
- Evetts, I. W. (1998) Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3): 198–202.
- Evetts, I. W., and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sunderland, MA: Sinauer Associates.
- Foreman, L. A., Champod, C., Evetts, I. W., Lambert, J. A., and Pope, S. (2003) Interpreting DNA evidence: A review. *International Statistics Journal*, 71: 473–473
- French, J. P. & Harrison, P. (2007) Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law*, 14(1): 137–144

- Friedman, R.D. (1996) Assessing evidence. *Michigan Law Review*, 94: 1810–1838.
- González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., and Ortega-García, J. (2006) Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2-3): 331–355.
- González-Rodríguez, J., Rose, P., Ramos, D., Torre, D., and Ortega-García, J. (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7): 2104–2115.
- Good, I. J. (1991) Weight of evidence and the Bayesian likelihood ratio. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 85–106. Chichester, UK: Ellis Horwood.
- Haigh, J. (2005) Review of statistics and the evaluation of evidence for forensic scientists. *Significance*, March: 40.
- Hodgson, D. (2002) A Lawyer looks at Bayes' Theorem. *The Australian Law Journal*, 76: 109–118.
- Hudson, T., de Jong, G., McDougall, K., Harrison, P., Nolan, F. (2007) F0 statistics for 100 young male speakers of standard Southern British English. In J. Trouvain and W. J. Barry (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences* 1809–1811.
- Jessen, M. (2008) Forensic phonetics. *Language and Linguistics Compass*, 2(4): 671–711.
- Lindley, D. V. (1991) Probability. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 27–50. Chichester, UK: Ellis Horwood.
- Lucy, D. (2005) *Introduction to Statistics for Forensic Scientists*. Chichester, UK: John Wiley.
- Nolan, F. (1983). *The phonetic Bases of Speaker Recognition*. Cambridge, UK: Cambridge University Press.
- Nolan, F. (1996) Forensic phonetics. Notes distributed at the two-week course at the 1996 Australian Linguistics Institute, Australian National University, Canberra.
- Nolan, F. (1997) Speaker recognition and forensic phonetics. In W. J. Hardcastle and J. Laver (eds) *The Handbook of Phonetic Sciences* 744–767. Oxford, UK: Blackwell.
- Nolan F., McDougall, K de Jong, G., and Hudson, T. (2006) A Forensic Phonetic Study of 'Dynamic' Sources of Variability in Speech: The DyViS Project. In Warren & Watson (eds.) *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* 13–18.
- Pigeon, S., Druyts, P., and Verlinde, P. (2000) Applying logistic regression to the fusion of the NIST '99 1-speaker submissions. *Digital Signal Processing*, 10: 237–248.

- Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10: 19–24.
- Robertson, B. and Vignaux, G.A. (1995) *Interpreting Evidence*. Chichester, UK: Wiley.
- Rose P (2002) *Forensic Speaker Identification*. London & New York: Taylor and Francis.
- Rose P (2003) *The Technical Comparison of Forensic Voice Samples*. Issue 99, *Expert Evidence*. Freckelton I, Selby H, (series eds.). Sydney, Australia: Thomson Lawbook Company.
- Rose (2005) Forensic Speaker Recognition at the beginning of the Twenty-First century. An overview and a demonstration. *Australian Journal of Forensic Sciences*, 37(2): 49–71.
- Saks, M. J., and Koehler, J. J. (2005) The coming paradigm shift in forensic identification science. *Science*, 309: 892–895.
- Thompson, W. C., and Schumann, E. L. (1987) Interpretation of statistical evidence in criminal trials. The prosecutor's fallacy and the defence attorney's fallacy. *Law and Human Behaviour*, 11: 167-187.

Una respuesta a la declaración de posicionamiento británico sobre la comparación forense de locutores

Phil Rose y Geoffrey-Stewart Morrison

School of Language Studies, Australian National University
[Escuela de Estudios Lingüísticos, Universidad Nacional de Australia]

Traducción por José Juan Lucena Molina

1. Introducción.

Un reciente número del *International Journal of Speech, Language and the Law* [Revista Internacional de Habla, Lenguaje y Ley] contenía una “declaración de posicionamiento sobre el uso de términos probabilísticos intuitivos en casos de comparación forense de locutores” (French y Harrison, 2007). Esta declaración de posicionamiento fue el resultado de un trabajo de colaboración llevado a cabo por una serie de investigadores y expertos forenses que trabajan en el Reino Unido. El prefacio establece que:

“la declaración circuló por todos los expertos de habla forenses y profesores universitarios interesados en el seno del Reino Unido. Con una única excepción, todos los que fueron consultados dieron su aprobación. La declaración refleja el posicionamiento casi unánime dentro del Reino Unido” (página 138).

La declaración fue entregada también a las Fiscalías de Escocia, Irlanda del Norte y de Inglaterra y Gales. Por simplicidad nos referiremos a ella como la *Declaración de posicionamiento del Reino Unido*, con la salvedad de que puede que no refleje el punto de vista de todas las partes interesadas en el Reino Unido¹, o que no tenga fuerza de ley en alguna jurisdicción dentro del Reino Unido.

Los editores del *International Journal of Speech, Language and the Law* invitaron a que hubiera respuestas a la *Declaración de posicionamiento del Reino Unido* en siguientes ediciones. La presente es nuestra respuesta (una versión preliminar fue presentada en julio de 2008 en el 17 congreso de la *International Association for Forensic Phonetics and Acoustics* [Asociación Internacional de Fonética y Acústica Forense]).

Primeramente resumimos la Declaración de posicionamiento del Reino Unido tal y como la entendemos y, más adelante, presentamos nuestra respuesta. Como ayuda para entender nuestra exposición precedemos nuestra respuesta con un resumen de lo que consideramos como el marco correcto de la presentación de la evidencia de comparación forense de voces.

2. Descripción de la Declaración de posicionamiento del Reino Unido.

Queda claro a resultas de lo expuesto en el prefacio de la Declaración de posicionamiento del Reino Unido que se refiere a la comparación de grabaciones de

¹ Puesto que conocemos que dos profesores universitarios interesados en el Reino Unido no fueron consultados, está claro que la *Declaración de posicionamiento del Reino Unido* no representa “el punto de vista de todos los expertos forenses de habla y profesores universitarios interesados en el Reino Unido” excepto uno.

voces realizadas por expertos y, de este modo, se relaciona con comparaciones de voces forense *technical* (técnicas) y no *naïve* (simplistas) (véase Nolan 1983: 7, 1997: 744-5).

2.1 Motivaciones y fines.

En su prefacio, la Declaración de posicionamiento del Reino Unido dice que fue motivada por una reflexión sobre “el marco en el que ordinariamente se formulan las conclusiones en casos de comparación forense de locutores” (página 137). La conciencia de que existe un problema en el actual marco se dice que se suscitó inicialmente a raíz del fallo en el caso R. contra Doheny y Adams ([1996] EWCA Crim 728) del Tribunal de Apelación de Inglaterra y Gales que asoció la falacia del Fiscal con la evidencia principal de un experto de ADN.

El prefacio de la Declaración de posicionamiento del Reino Unido sostiene que presenta:

“... [un] nuevo enfoque [el cual] provoca un cambio fundamental en el papel del analista y la evidencia. En el pasado, los expertos forenses de habla pasaron por ser identificadores de locutores. Dentro del nuevo enfoque ellos no realizan identificaciones. En su lugar, su papel consiste en proporcionar una valoración sobre si la voz en la grabación cuestionada se corresponde con la descripción del sospechoso.” (página 138)

La nota al pie de página nº 2 del Documento de posicionamiento del Reino Unido añade que la actividad pasa a ser considerada como comparación, no como identificación. El prefacio también sostiene que el fin de llevar a cabo la Declaración de posicionamiento del Reino Unido fue “... llevar la disciplina [de comparación forense de voz] dentro del pensamiento moderno en otras áreas de la ciencia forense” (página 137), y que “Este nuevo marco es, en el nivel conceptual, idéntico al que se utiliza hoy día en la presentación de la evidencia de ADN” (página 138).

Al final del documento, los autores y los firmantes del Documento de posicionamiento del Reino Unido reconocen que ellos:

“... aceptan en principio el deseo de enmarcar el trabajo de la comparación de locutores forense en un marco conceptual (incluyendo que sea bayesiano) de una relación de verosimilitud. Sin embargo, [ellos] consideran que la falta de datos demográficos junto con los problemas de definir las poblaciones de referencia relevantes son razones que impiden la aplicación cuantitativa de este tipo de enfoque en el presente contexto.” (página 142, §6).

Parece, pues, que la renuncia racional de la Declaración de posicionamiento del Reino Unido del marco basado en la relación de verosimilitud a favor de una apuesta alternativa se vislumbra como un mal menor. Sin embargo, esto sería una caracterización injustificada. Se han realizado claros intentos para conseguir una solución con enfoques basados en relaciones de verosimilitud, y podemos ver en esta solución – aunque no sea explícitamente llamada así – una motivación adicional para el fin perseguido. Lo que argumentaremos es que no es la mejor solución, y que

no es conceptualmente equivalente al marco en el que se presenta la evidencia de ADN.

2.2 El marco de la Declaración de posicionamiento del Reino Unido.

En la figura 1 presentamos un diagrama de flujo del marco propuesto en la Declaración de posicionamiento del Reino Unido. En este marco, las muestras de habla ha de compararse en términos de dos factores ordenados en serie: *coherencia* y *peculiaridad*.

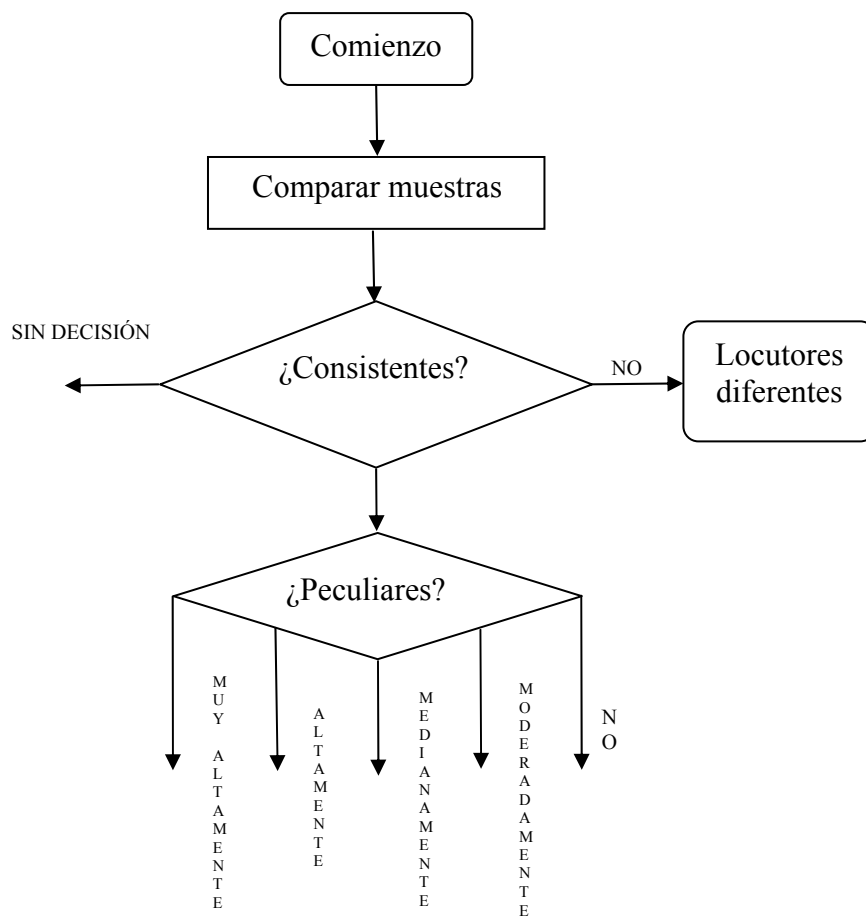


Figura 1. Representación del diagrama de flujo del marco de la Declaración de posicionamiento del Reino Unido.

2.2.1 Consistencia.

La consistencia se caracteriza por lo siguiente: “si las muestras cuestionada y conocida son compatibles, o consistentes, con el hecho de haber sido producidas por el mismo locutor” (página 141, §4.1). Se valora por “el grado de similitud o diferencia entre las propiedades observables” (página 141, §4.1). Las diferencias entre las muestras cuestionada y conocida van en contra de la consistencia a menos que “puedan explicarse por modelos de variación acústica, fonética o lingüística (por ejemplo, por referencia a las distintas características de canal, o por factores [intra-locutor] sociolingüísticos,

psicológicos y/o físicos)” (página 141, §4.1). La consistencia se cuantifica en una escala de tres niveles: *consistente*, *no consistente*, o *sin decisión*.

Si las muestras *no son consistentes*, se considera que han sido pronunciadas por distintos locutores. Si son *consistentes*, se procede a considerar la cuestión de la peculiaridad (de este modo, consistencia y peculiaridad están ordenados en serie, siendo sólo posible un juicio sobre la peculiaridad en el supuesto de que primeramente se haya determinado positivamente la consistencia). No existen instrucciones sobre qué acciones realizar cuando se determine la opción *sin decisión*.

2.2.2 Peculiaridad.

La Declaración de posicionamiento del Reino Unido enfatiza que una determinación positiva de consistencia no implica necesariamente que las muestras cuestionada y conocida fueran pronunciadas por la misma persona, puesto que “el conjunto de propiedades que conducen a la decisión de consistencia ... [pudieran] estar compartidas por un número sustancial de personas distintas entre la población” (página 141, §4.2). La Declaración de posicionamiento del Reino Unido implica que la probabilidad de que las muestras hayan sido pronunciadas por las mismas personas será mayor si el conjunto de propiedades compartidas es peculiar o inusual. La peculiaridad se valora en una escala de cinco niveles desde la *no peculiaridad* a una *peculiaridad muy alta*, siendo esta última interpretada como que “la posibilidad de que esta combinación de propiedades sea compartida por varios locutores es muy remota” (página 141, §4.2). No existen instrucciones sobre el modo de proceder cuando se toma una decisión sobre la peculiaridad. Presumimos que el experto deberá informar sobre si las muestras son consistentes con haber sido pronunciadas por una misma persona, y proporcionar el grado determinado de peculiaridad como indicador de cuánto de inusual es haber encontrado esta consistencia si las dos muestras no hubieran sido pronunciadas por la misma persona.

3. Respuesta a la Declaración de posicionamiento del Reino Unido.

Primeramente debemos decir que aplaudimos la motivación que guía a la Declaración sobre posicionamiento del Reino Unido, y damos la bienvenida a su orientación general. Específicamente, estamos de acuerdo completamente con los objetivos de llevar la presentación de la evidencia de comparación de voz forense en la línea del pensamiento moderno en otras áreas de la ciencia forense, particularmente en llevarla en la línea de la práctica moderna de la evaluación de ADN. Estos objetivos no son nuevos: fueron también propuestos por Champod y Meuwly (2000), González-Rodríguez y otros (2006), González-Rodríguez y otros (2007), Rose (2002, 2003) y Saks y Koehler (2005). Argumentaremos, sin embargo, que la Declaración de posicionamiento del Reino Unido falla en la búsqueda de esos objetivos.

Antes de realizar un análisis crítico de la Declaración de posicionamiento del Reino Unido, haremos la discusión más fácil si primeramente presentamos lo que consideramos el marco correcto, lógica y legalmente considerado, para la evaluación de la evidencia de comparación forense: el marco de la relación de verosimilitud. Este es el marco que creemos que representa el moderno pensamiento en otras áreas de la ciencia forense como el ejemplificado por la práctica actual de evaluación de la evidencia de ADN. Esta posición está apoyada por numerosos libros de texto, artículos y críticas

escritos por estadísticos forenses, expertos en leyes, y científicos forenses, como por ejemplo Aitken y Taroni (2004), Balding (2005), Champod y Meuwly (2000), Evett (1991, 1998), Friedman (1996), Good (1991), González-Rodríguez y otros (2006), Haigh (2005), Hodgson (2002), Lindley (1991) y Robertson y Vignaux (1995).

3.1 El marco de la relación de verosimilitud.

Lo que contamos a continuación está extractado de Rose (2005: 49-54). Está escrito desde la perspectiva de que los objetos comparados sean muestras de habla, pero es, en principio, aplicable a la evaluación de cualquier tipo de evidencia forense (ADN, huellas dactilares, balística, trazas de herramienta, etc...) donde se comparan muestras dubitada e indubitada y es posible cuantificar las propiedades físicas que pueden variar de una muestra a otra. Para los interesados en una más completa exposición del marco de la relación de verosimilitud, recomendamos los siguientes textos: Balding (2005), Lucy (2005), Robertson y Vignaux (1995) y Rose (2002, 2003).

En una comparación de voz forense, se compara típicamente una grabación de voz desconocida, generalmente la de un criminal, con una o más grabaciones de una voz conocida, generalmente de un sospechoso o imputado. Las partes interesadas (policía/fiscalía, Jurado o Tribunal, abogados) quieren conocer si la voz desconocida (o cuestionada) procede del mismo locutor que la voz conocida. Generalmente comprenden que no es posible dar una respuesta definitiva; un juicio es, después de todo, una toma de decisiones ante la incertidumbre. Así, ellos normalmente preguntarán: *¿ cuál es la probabilidad de que las muestras hayan sido pronunciadas por la misma persona ?* – una muy razonable forma de hacerlo puesto que tanto filósofos como estadísticos están de acuerdo en que la mejor forma de cuantificar la incertidumbre es mediante probabilidades (Lindley, 1991). Implicado, desde luego, está también el papel que desempeña la evidencia. Es decir, la pregunta debería ser realmente así: *¿ cuál es la probabilidad, dada la evidencia de voz, de que las muestras hayan sido pronunciadas por la misma persona ?*. Esto se formaliza, convencional y convenientemente, por la expresión de probabilidad condicional siguiente:

$$P(H_{ml} | E_{hb}) \quad (1)$$

donde “p” significa probabilidad, “H_{ss}” la hipótesis (**h**ypothesis) del Fiscal consistente en que está involucrado el **mismo locutor**, la barra vertical significa “dada/o” o “condicionada/o a”, y “E_{hb}” significa evidencia de habla – la inevitable presencia de diferencias entre las muestras de habla del criminal y del sospechoso -. Es corriente utilizar apuestas (odds) en lugar de probabilidades, así la representación formal de la ecuación se convierte en:

$$P(H_{ml} | E_{hb}) / P(H_{dl} | E_{hb}) \quad (2)$$

donde “H_{dl}” significa la hipótesis de que las muestras fueran pronunciadas por **distintos locutores**.

La solución de esta ecuación viene dada por el Teorema de Bayes, el cual es conocido, al menos, desde la mitad del siglo XVIII (Bayes, 1763). El Teorema de Bayes es de suma importancia cuando se quiere conocer la probabilidad de una hipótesis dada la

evidencia, y esto es lo que produce su especial relevancia en identificación forense. El Teorema de Bayes establece, informalmente, que la probabilidad de la hipótesis, dada la evidencia, puede estimarse a partir de dos cosas: (1) qué probabilidad tiene la hipótesis antes de conocer la evidencia; y (2) la fuerza de la evidencia.

La forma de apuesta del Teorema de Bayes, aplicada a la comparación de voz forense, viene dada por la ecuación (3). Se dice que la apuesta a favor de que las muestras procedan del mismo locutor, dada la evidencia de habla (esto es lo que todo el mundo desea saber y que se denomina apuesta a posteriori, y que se encuentra a la izquierda de la igualdad), se calcula multiplicando la apuesta a priori a favor de que las muestras procedan del mismo locutor por la fuerza de la evidencia. De este modo, la apuesta a favor de que las muestras procedan del mismo locutor puede calcularse con dos términos: *la apuesta a priori* y *la relación de verosimilitud*.

$$P(H_{ml} | E_{hb}) / P(H_{dl} | E_{hb}) = [P(H_{ml}) / P(H_{dl})] * [P(E_{hb} | H_{ml}) / P(E_{hb} | H_{dl})] \quad (3)$$

Apuesta a posteriori = Apuesta a priori * Relación de verosimilitud

La *apuesta a priori* es la apuesta a favor de la hipótesis antes de que la evidencia de voz sea requerida. Se trata, simplemente, de la probabilidad de que las muestras procedan del mismo locutor dividido por la probabilidad de que las muestras procedan de locutores distintos. En el límite, podría ser de algún habitante del planeta, pero generalmente el margen es considerablemente más estrecho al tener en cuenta información obvia en la voz como el sexo y el acento, así como cualquier otro tipo de información pragmática.

La *relación de verosimilitud* es la métrica más importante en comparación de voz forense porque es una medida de la *fuerza de la evidencia* a favor de una hipótesis, y es lo que el experto debería intentar estimar. La fórmula (3) muestra que la relación de verosimilitudes es también una relación de probabilidades, pero esas probabilidades son probabilidades de la *evidencia*, no de las *hipótesis*. La relación de verosimilitud cuantifica cuánto más probable es la obtención de las diferencias entre las muestras de habla del sospechoso y del criminal asumiendo que procedan del mismo locutor respecto a que procedan de locutores diferentes.

Si resultara más probable obtener la evidencia de habla si las muestras procedieran del mismo locutor que de locutores diferentes – si $P(E_{hb} | H_{ml})$ es mayor que $P(E_{hb} | H_{dl})$ – se apoyaría la hipótesis del Fiscal consistente en que las muestras proceden del mismo locutor. Si, por el contrario, resultara más probable obtener la evidencia de habla si las muestras procedieran de locutores distintos que del mismo locutor – si $P(E_{hb} | H_{dl})$ es mayor que $P(E_{hb} | H_{ml})$ – se apoyaría la hipótesis de la defensa consistente en que las muestras proceden de locutores diferentes. Si fuera tan probable obtener la evidencia asumiendo que las muestras proceden del mismo locutor como de locutores diferentes – si la relación $P(E_{hb} | H_{ml}) / P(E_{hb} | H_{dl})$ es igual a la unidad – la evidencia no proporciona información útil.

De este modo, la magnitud de la relación de verosimilitud cuantifica la fuerza de la evidencia: si es mayor a la unidad, se apoya que las muestras proceden del mismo locutor; si es menor a la unidad, se apoya que las muestras proceden de locutores

distintos; si vale la unidad (o valores próximos a ella), la evidencia no ofrece información (o casi no la ofrece).

Los principales libros de texto sobre evaluación de la evidencia forense (por ejemplo, Robertson y Vignaux, 1995) y sobre estadística forense (por ejemplo, Aitken y Stoney, 1991; Aitken y Taroni, 2004; Lucy, 2005), subrayan que le corresponde al experto forense cuantificar la fuerza de la evidencia estimando la relación de verosimilitud: las probabilidades de la evidencia bajo las hipótesis competitivas del Fiscal y de la defensa:

“Este enfoque, tanto si el asunto en cuestión es una prueba de ADN, de fragmentos de cristal, de fibras o cualquier otra, es aplastante” (Haigh, 2005).

“La evaluación estadística y, particularmente, los métodos bayesianos como el cálculo de relaciones de verosimilitud ... son el único medio racional demostrable de cuantificar el valor de la evidencia actualmente disponible: todo lo demás es intuición o conjetura.” (Lucy, 2005: 138).

3.2 Comparación, no identificación.

La Declaración de posicionamiento del Reino Unido enfatiza que su propuesta no se considera una identificación, sino una comparación, y éste es el término que ellos usan (y que nosotros hemos adoptado). Esta elección vale la pena comentarla porque existen muchos términos actualmente en uso: *identificación*, *reconocimiento*, *verificación* y *discriminación*. En la literatura puede observarse que algunos de esos términos, por ejemplo *identificación* y *reconocimiento*, se han utilizado indistintamente, y que términos distintos, por ejemplo *identificación* y *verificación*, han sido lógicamente utilizados para designar distintos tipos de análisis o diferentes aplicaciones (Rose, 2002: Capítulo 3). Estamos de acuerdo con la Declaración de posicionamiento del Reino Unido en que el término *comparación* es el más apropiado, pero por razones ligeramente diferentes. Términos como *identificación*, *verificación* y *reconocimiento* implican la expresión de una probabilidad a posteriori (es decir, la probabilidad de que las voces del sospechoso y del criminal procedan de la misma persona), y algunas tienen connotaciones de proporcionar decisiones categóricas. Argumentamos más abajo, como se dijo en Rose (2002: 89), que ya que no es posible desde el punto de vista de la lógica, y a que resulta inapropiado desde el punto de vista legal, que un experto proporcione probabilidades a posteriori, no hay posibilidad de que realice una identificación, una verificación o un reconocimiento y, por consiguiente, estamos de acuerdo con la Declaración de posicionamiento del Reino Unido que, con respecto al trabajo forense, esos términos deberían evitarse a favor de un término neutral como el de *comparación*, que no tiene connotaciones de proporcionar una decisión en forma de probabilidades a posteriori. También sugerimos que como los objetos comparados son grabaciones de voces – las voces son las que se comparan, no los locutores – es mejor adoptar el término *comparación forense de voces* en lugar de *comparación forense de locutores* tal y como se recoge en la Declaración de posicionamiento del Reino Unido. Nolan (1983, 1996) aporta la mejor caracterización actual de una voz con fines forenses. Su versión semiótica se describe en detalle en Rose (2002: Capítulo 10).

3.3 Prohibición sobre probabilidad de la hipótesis, dada la evidencia.

La primera propuesta importante de la Declaración de posicionamiento del Reino Unido es la recomendación de que el experto refrane calcular probabilidades de las hipótesis, dada la evidencia [$p(H|E)$]. Apoyamos esto fuertemente: desde hace algún tiempo hasta la fecha ésta ha sido la posición adoptada por los estadísticos forenses y, más recientemente, por algunos Tribunales (consúltese una discusión histórica sobre este tema en Aitken y Taroni 2004: 108, 122-128, 153-155, 208-213; Balding 2005: 145-153). Sin embargo, la Declaración de posicionamiento del Reino Unido dice que la razón por la que un experto forense no debe citar $p(H|E)$ es porque esta probabilidad aporta una “ponderación falsa” a la evidencia que se relaciona con la falacia del Fiscal. La falacia del Fiscal consiste en una transposición errónea de la evidencia con la hipótesis, es decir, se reemplaza $p(E|H)$ por $p(H|E)$. Esto es lo mismo que decir que como la evidencia es 1000 veces más probable bajo la asunción de culpabilidad, el imputado es 1000 veces más probable de ser culpable (Aitken y Taroni 2004: 79-82; Balding 2005: 146-147; Donnelly 2005; Evett 1998, Thompson y Schumann, 1987). Ciertamente, los expertos forenses deben evitar cometer ellos mismos este tipo de errores, y deben realizar todo lo posible dentro de sus atribuciones para que no los cometan los abogados, los Jueces y los Jurados. Pero los argumentos de la Declaración de posicionamiento del Reino Unido sobre “ponderación falsa” no se dirige a las auténticas razones sobre por qué los expertos forenses deben proporcionar la probabilidad de la evidencia, dada la hipótesis [$p(E|H)$], y por qué no pueden proporcionar la probabilidad de la hipótesis, dada la evidencia [$p(H|E)$].

Hay dos razones por las que un experto forense no puede proporcionar la probabilidad de la hipótesis, dada la evidencia: una lógica, y otra legal. La razón lógica se deriva trivialmente del Teorema de Bayes. La apuesta a posteriori se determina por dos cosas: la fuerza de la evidencia (relación de verosimilitud) y la apuesta a priori (consúltese la ecuación 3). Al experto no le incumbe la apuesta a priori, por lo que tampoco le compete, lógicamente, la apuesta a posteriori. La razón legal tiene que ver con la violación de la regla de la cuestión última: en los casos en los que la muestra procedente del criminal sea verdaderamente incriminante, el pronunciamiento del experto sobre si el sospechoso es quien probablemente haya pronunciado el habla incriminante es equivalente a una expresión de probable culpabilidad y, de ese modo, usurpa el papel de quien la potestad para juzgar los hechos (el Juez o el Jurado, dependiendo del sistema legal).

Aunque la Declaración de posicionamiento del Reino Unido aparentemente condena la práctica de proporcionar probabilidades de hipótesis, dada la evidencia, hay dos momentos en los que se recomienda, de hecho, proporcionar $P(H|E)$. Estos se discuten en las siguientes dos subsecciones, una relacionada con las diferencias entre los datos de ADN y los de habla, y el otro con comparaciones en conjunto cerrado.

3.3.1 Diferencias entre los datos de ADN y los de habla.

La primera violación de la prohibición de proporcionar $p(H|E)$ ocurre en relación con la determinación de no consistente en la cuestión de la consistencia:

“Cuando las muestras no son consistentes, no vemos un error lógico si se concluye que las muestras proceden de personas diferentes. Esto puede sostenerse con un grado de confianza apropiado a las exigencias de los datos.” (página 141, §4.3).

Decir, con una confianza dada, que las muestras proceden de personas diferentes porque no son consistentes es pronunciarse sobre $p(H|E)$. De acuerdo con lo que la Declaración de posicionamiento del Reino Unido sostiene más arriba, aplicando el Teorema de Bayes eso es, de hecho, un error lógico.

Sospechamos que esta inconsistencia se ha producido porque los autores de la Declaración de posicionamiento del Reino Unido intentaron adaptar un modelo para el análisis de ADN sin tener en cuenta importantes diferencias en la naturaleza del ADN respecto a la evidencia de habla. Aunque la evaluación de la evidencia de habla forense puede ciertamente hacerse del mismo modo que la de ADN, con relaciones de verosimilitud – esto fue demostrado en una reciente publicación con enfoques automático y tradicional (González-Rodríguez y otros, 2007) – se ha de tener cuidado cuando se hagan comparaciones entre los datos de ADN y los de voz forenses. Esto se debe a las diferencias en la naturaleza de la variación en cada caso. Los tres aspectos de variación más importantes en las ciencias forenses son el **tipo** de variación; cuántos **niveles** de variación existen; y la **magnitud** de la variación. Los datos de ADN difieren de los datos de voz en los tres, pero los que son importantes en este caso son el tipo y los niveles de variación.

Las variables pueden ser continuas o discretas, o una combinación de esos tipos. Las variables de ADN – típicamente la localización de los alelos STR en determinados lugares de la cadena de ADN – son discretas. Con variables discretas es posible hablar de plena coincidencia, por ejemplo que ambas muestras tengan un genotipo con la misma combinación alélica: 14, 16 en el locus D18, y 9.3, 9.3 en el locus THO1 (Balding 2005: 3) [9.3, aunque parezca continuo, no lo es; significa que una de las repeticiones sólo tiene tres en lugar de las cuatro esperadas bases comunes]. En ADN también es posible una ausencia plena de coincidencia.

Mientras que el ADN es discreto, la evidencia de habla es continua: los coeficientes cepstrales, las frecuencias centrales formánticas, etc..., son variables continuas, e incluso las propiedades de alto nivel como la incidencia de un particular alófono se valoran con proporciones continuas. También, mientras que las propiedades del habla varían de un momento a otro, el ADN de un organismo biológico será siempre el mismo cada vez que sea medido (salvo errores de medida, contaminación, cambios somáticos, trasplantes, quimeras, etc...).

Salvo excepciones, entonces, las propiedades de categoricidad e invariabilidad conllevan que si dos perfiles de ADN no coinciden, la probabilidad de que ocurra semejante cosa asumiendo que proceden del mismo organismo es cero. En este caso, el numerador de la relación de verosimilitud es cero y la probabilidad a posteriori de que vengan del mismo organismo, con independencia de los prioris, es también cero. Del mismo modo, el ADN puede utilizarse para proporcionar evidencias definitivas de exclusión. No así con el habla. En general, los datos de habla no permiten, por su naturaleza, tal clase de exclusión definitiva. Podemos imaginar algunas condiciones bajo las cuales una comparación de voces pudiera concluir con una exclusión definitiva, por ejemplo el tracto vocal de un niño no podría generar los formantes más bajos de un

adulto varón típico, pero en tales casos las voces sonarían tan distintas que sería muy improbable que se solicitara un informe forense a un experto.

De nuevo, salvo excepciones, dada una plena coincidencia entre dos perfiles de ADN, la probabilidad de observar semejante hecho asumiendo que ambas muestras proceden del mismo organismo es uno (Aitken y Taroni 2004: 404, Evett, 1998). Siendo el numerador de la relación de verosimilitud igual a la unidad, su magnitud es dependiente del tamaño de su denominador. El denominador es la *random match probability* [probabilidad de coincidencia aleatoria] (referida en la Declaración de posicionamiento del Reino Unido como *random occurrence ratio* [tasa de ocurrencia aleatoria²]). Se trata de la probabilidad de observar una coincidencia aleatoria con el perfil de ADN obtenido por parte de alguno de los perfiles de los miembros de una población relevante. Como bajo estas circunstancias la relación de verosimilitud es equivalente a la inversa de la probabilidad de coincidencia aleatoria, la fuerza de la evidencia de ADN puede presentarse también en la forma de probabilidad de coincidencia aleatoria en lugar de una relación de verosimilitud.

Es posible que una inapropiada transferencia del análisis de ADN en la Declaración de posicionamiento del Reino Unido, haya producido también un problema con el concepto de coincidencia aleatoria. La Declaración de posicionamiento del Reino Unido considera un caso, en dicho lugar, en el que se ha producido una coincidencia de los perfiles de ADN del sospechoso y del criminal, con una probabilidad de coincidencia aleatoria de 1 entre 1 millón (es decir, una persona entre un millón tiene un perfil coincidente con el del criminal), teniendo en cuenta que en el Reino Unido viven 60 millones de personas. Bajo esas circunstancias, la Declaración de posicionamiento del Reino Unido dice: "... hay una probabilidad de 1 / 60 de que el ADN proceda del imputado" (página 139)³. La Declaración de posicionamiento del Reino Unido continúa:

"La estimación de que una persona entre 1 millón comparte el perfil de ADN se conoce como 'tasa de ocurrencia aleatoria'. Los fonetistas pueden calcular la tasa de ocurrencia aleatoria de muy pocas propiedades del habla. Resultan una excepción la frecuencia fundamental (una medida del tono de la voz), la tasa de articulación (velocidad de habla) y el tartamudeo" (página 140, §3).

Como los datos de habla son inherentemente continuos y es trivial que un locutor nunca puede decir exactamente la misma cosa dos veces, hay siempre variación entre las muestras de habla, por lo que el numerador de la relación de verosimilitud derivado de una comparación de voz forense nunca podrá ser cero o uno. El concepto de coincidencia aleatoria es, por consiguiente, no aplicable a los datos de habla continuos.

² El término '*random occurrence ratio*' [tasa de ocurrencia aleatoria] empleado por el Tribunal [en R. contra Dohney y Adams] aparenta ser un sinónimo de probabilidad de coincidencia. Este nuevo término inventado es una adición incómoda a los muchos términos ya disponibles: la ausencia de familiaridad con ella puede llevar a confusión". Balding (2005: 152).

³ La respuesta correcta es una probabilidad de 1 / 61 de que el imputado sea la fuente de la traza de ADN (o una apuesta de 1 contra 60). De los 60 millones de los posibles perpetradores del crimen en el Reino Unido, uno es culpable y el resto, 59.999.999, son inocentes. El culpable producirá una coincidencia de perfiles, y entre los restantes 59.999.999, se producirán 60 coincidencias de perfiles (porque la probabilidad de una coincidencia aleatoria es 1 / 1.000.000, y 59.999.999 x (1/1.000.000) = 60 (redondeando al entero más próximo). Así, habrá un total de 61 posibles coincidencias. De las 61 coincidencias, sólo 1 es la coincidencia verdadera, y el resto son falsos positivos, por tanto, la probabilidad de que el sospechoso fuera el autor de la traza es de 1 / 61. En Balding (2005: 11) puede encontrarse una fórmula simplificada para calcular la probabilidad de culpabilidad bajo estas circunstancias [$P(G|E) = 1 / 1+N*p$], donde $P(G|E)$ significa probabilidad de culpabilidad (Guilt) dada la evidencia (Evidence), N es el número de personas distintas al sospechoso que pudieran haber sido autores del crimen, y p es la probabilidad de una coincidencia aleatoria.

La fuerza de la evidencia de una comparación forense de voz sólo puede expresarse en la forma de una relación de verosimilitud.

Mientras que la probabilidad de coincidencia aleatoria es, ciertamente, un concepto sin significado con respecto a propiedades evaluadas inherentemente de forma continua como la frecuencia fundamental, podría argumentarse que podrían calcularse probabilidades de coincidencia aleatoria en incidencias de propiedades del habla como el tartamudeo. Sin embargo, esto sólo podría hacerse bajo las asunciones de que en una grabación de alguien que habitualmente tartamudee siempre habrá pasajes de tartamudeo, y de que en una grabación de alguien que no lo haga generalmente, nunca habrá pasajes de tartamudeo.

3.3.2 Comparaciones en conjunto cerrado.

La segunda violación de la prohibición de aportar la probabilidad $p(H|E)$ ocurre en la sección 5 de la Declaración de posicionamiento del Reino Unido:

“En muy pocos casos, sin embargo, existe evidencia independiente (por ejemplo, videovigilancia) para sólo un conjunto cerrado de locutores que estuvieron presentes e intervinieron en la conversación. En tales casos, la tarea de la comparación se convierte en saber quién dijo qué. En esas circunstancias, si las voces son suficientemente diferentes unas de otras, consideramos justificado que puedan formularse pronunciamientos categóricos sobre identificación” (página 142, §5).

Realizar manifestaciones categóricas de identificación, dada una suficiente distinción entre lo que se estudia, es una afirmación sobre la probabilidad de la hipótesis dada la evidencia y una violación de naturaleza lógica del Teorema de Bayes. Las comparaciones en conjunto cerrado pueden tratarse de igual forma que las de conjunto abierto desde el punto de vista de la fuerza de la evidencia (consúltese Rose 2002: 64, 74).

3.4 Valoración en dos etapas.

Otra parte importante de la propuesta, y de nuevo un paso en la dirección correcta, es la valoración bipartita entre consistencia y peculiaridad. La Sección 4.2 de la Declaración de posicionamiento del Reino Unido acierta asumiendo que el valor de la evidencia no sólo depende de la similitud entre las dos muestras, sino también de su peculiaridad. Dos muestras muy similares, pero muy poco peculiares, no serán evaluadas con un valor alto desde el punto de vista de la fuerza de la evidencia a favor de la identidad que en el caso de dos muestras muy similares pero muy atípicas. Este es un asunto no siempre comprendido, y no es poco común encontrar que la asunción de identidad descansa sólo en la similitud. Resulta, por consiguiente, satisfactorio ver este extremo aclarado en la Declaración de posicionamiento del Reino Unido.

A primera vista, los términos de consistencia y peculiaridad parecen paralelos al numerador y denominador de una relación de verosimilitud como la tratada en la sección 3.1. Sin embargo, el uso por parte del marco de la Declaración de una valoración bipartita a través de la consistencia y de la peculiaridad no es equivalente al

cálculo de la relación de verosimilitud. Una propiedad esencial de la relación de la verosimilitud es que el numerador y el denominador se miden en la misma escala (son ambos valores de densidades de probabilidad) y están directamente asociados el uno con el otro (es decir, en la forma de una relación). En el marco de la Declaración, la consistencia y la peculiaridad están ordenadas en serie; se miden en distintas escalas (la primera tiene tres niveles y la segunda cinco); y no están directamente relacionadas entre sí.

El Tribunal o Jurado necesita saber si las diferencias entre muestras de habla son más probables si fueron pronunciadas por un mismo locutor que si lo fueron por locutores distintos, o si son equiprobables en ambos casos. No es posible hacer esto si ambos términos no se cuantifican en la misma escala y están directamente relacionados entre sí. Consideramos esto como un punto débil del marco de la Declaración de posicionamiento del Reino Unido.

El análisis en dos etapas del marco de la Declaración sobre consistencia y peculiaridad es, de hecho, reminiscencia de la evaluación de la evidencia de Evett (1977) en términos de etapas de comparación y de significación. En ese enfoque, primero uno decide si hay una coincidencia sobre la base de los criterios acordados previamente – digamos que las muestras se encuentren dentro de tres desviaciones estándar una de otra -. Después, se valora la probabilidad de encontrar el grado observado de similitud en la población relevante (véanse críticas de este marco en Aitken y Taroni, 2004: 10-11, y en Evett, 1991). Aunque se han aplicado, históricamente, algunas variantes de esta valoración en dos etapas a la evidencia de ADN, fue sustituida por la evaluación de la relación de verosimilitud, de modo que el enfoque de dos etapas no representa la práctica moderna (consúltese Foreman y otros, 2003, para encontrar un repaso histórico de la interpretación de la evidencia de ADN hasta esa fecha).

Además de los problemas sistemáticos con la valoración en dos etapas de la Declaración de posicionamiento del Reino Unido, existen problemas con la estructura de sus partes componentes. Más abajo discutimos el problema del efecto acantilado y el de los datos multivariantes, ambos relacionados con el hecho de que la consistencia y la peculiaridad en el marco de la Declaración tiene un número finito de salidas categóricas. Discutimos estos problemas en relación con la consistencia, pero debe quedar claro que existen argumentos idénticos con respecto a la peculiaridad. La división en tres frente a cinco salidas categóricas es irrelevante para la lógica argumental. También discutimos la elección desafortunada de la palabra “consistente”.

3.4.1 El efecto acantilado.

Es una trivialidad fonética decir que haya siempre diferencias entre muestras de habla, incluso aunque provengan de un mismo locutor que repita la misma cosa pocos segundos después de la primera vez. Además, esas diferencias son graduables, no categóricas, por lo que tales diferencias entre las muestras no pueden encajarse adecuadamente dentro de la terna de salidas categóricas del marco de la Declaración.

Consideremos los formantes vocálicos como propiedades graduables típicas. Se suscita la situación cuando se comparan un conjunto de vocales pronunciadas por el sospechoso y el criminal con respecto a sus distribuciones de frecuencias centrales formánticas. La Tabla 1 recoge los valores de media y desviación estándar calculados a partir de los

formantes de realizaciones de un único fonema vocálico extraídos de dos conversaciones telefónicas interceptadas. Para fijar el argumento asumamos que se ha determinado que las distribuciones de las propiedades son suficientemente próximas a la normal para garantizar que el uso de la media y de la desviación típica sean representaciones apropiadas de la tendencia central y de la dispersión.

Tabla 1. Medias y desviaciones estándar de frecuencias centrales formánticas medidas en Hz en 15 realizaciones de /ə:/ en inglés australiano en dos conversaciones telefónicas interceptadas distintas.		
Sospechoso	F2	F3
Media	1429	2298
Desviación estándar	30	67
Criminal		
Media	1450	2329
Desviación estándar	48	56
Diferencia entre las medias	21	31

Consideremos F2 como una propiedad singular. Imaginemos que, incluso en ausencia de datos poblaciones específicos, sobre la base de su experiencia, la mayor parte de los expertos estarían de acuerdo en que 21 Hz (una diferencia del 1.5%) estaría dentro del margen de diferencia esperado para la media de F2 de la vocal neutra de un mismo locutor hablando normalmente en distintas ocasiones. Si trabajamos en un marco como el de la Declaración, decidiríamos que existe consistencia. Como el marco de la Declaración no ha especificado cómo llegar a la declaración de consistencia, asumiremos una implícita referencia a alguna cuantificación de la variación como la desviación estándar de las propiedades. La Tabla 1 muestra que la desviación estándar del sospechoso y del criminal en F2 es mayor que la diferencia de medias en F2 entre las muestras. Pero, ¿en qué momento cambiamos la decisión sobre si los valores observados son consistentes o no con respecto a su procedencia de un mismo locutor ?. ¿ Debería estar el límite en dos desviaciones estándar ?, o ¿ quizá, en tres ?. ¿ Debemos aplicar un test estadístico frecuentista, como un t-test con un nivel de significación α del 0.05, o quizá del 0.01 ?. Tal enfoque impone una decisión categórica, con el correspondiente efecto acantilado (Robertson y Vignaux, 1995: 118). Si el límite se colocara en dos desviaciones estándar utilizando la desviación estándar de los datos del sospechoso, una diferencia en media de F2 de 59.9 Hz sería considerada *consistente*, pero una diferencia de 60.1 Hz sería considerada *no consistente* (¿ o en ambos casos sería declarada una *no decisión* ?). ¿ Puede una decisión depender de una diferencia de 0.2 Hz ?. Nosotros sostenemos que cualquier métrica de similitud que pueda usarse en una comparación forense de locutores debe ser gradual, no categórica.

3.4.2 Problemas con datos multivariantes.

Detectamos la presencia de otro problema en el factor de consistencia del marco de la Declaración cuando se necesitan comparar muestras con respecto a múltiples características. El prefacio de la Declaración de posicionamiento del Reino Unido describe el proceso de comparación de voz forense como:

“[conllevando] ‘segmentar’ las muestras en sus ‘aspectos’ acústicos y fonéticos constitutivos (por ejemplo, calidad de la voz, entonación, ritmo, tempo, tasa de articulación, realizaciones de vocales y consonantes) y analizando cada una aisladamente” (página 138).

No vemos ningún problema en ello puesto que la multidimensionalidad es uno de los factores que contribuye a la discriminabilidad forense de las voces. Sin embargo, la

Declaración no da indicación alguna sobre cómo combinar finalmente la evidencia individual de cada uno de esos ‘aspectos’. Volviendo a los datos formánticos de la Tabla 1 y al ejemplo del límite de dos veces la desviación estándar basado en la desviación estándar de los datos del sospechoso (el argumento se aplicaría igualmente bien cualesquiera que fueran las propiedades en consideración o el umbral prescrito), ¿cuál sería la decisión con respecto a la consistencia si la diferencia en las medias de F2 fuera de 50 Hz, dentro del límite, pero la diferencia en las medias de F3 fuera de 140 Hz, fuera del límite ?. ¿Qué decisión tomaríamos si un par de muestras se juzgaran consistentes en nueve propiedades pero no consistentes en una más ?. De nuevo, incumbe al marco de la Declaración especificar cómo deberíamos proceder en esas circunstancias. Sobre esto, como con otros problemas semejantes, un ejemplo de utilización del enfoque defendido sería útil, pues a nosotros nos resulta muy difícil llevarlo a la práctica. Si renunciamos a una cuantificación categórica, podemos utilizar simplemente una métrica gradual multivariante de similitud, tal como el numerador de la relación de verosimilitud multivariante. La literatura contiene ya muchos procedimientos que tratan con datos multivariantes dentro del marco de la relación de verosimilitud (por ejemplo, Aitken y Lucy, 2004; Pigeon, Druyts y Verlinde, 2000; Reynolds, Quatieri y Dunn, 2000).

3.4.3 La semántica de ‘consistente’.

La Sección 4.1 del marco de la Declaración define consistencia como “el grado en el que las propiedades observadas [son] similares o diferentes” (página 141, §4.1). Esta es, ciertamente, una noción coherente por la que es posible estimar cómo de similares son las voces cuestionada y conocida en las propiedades especificadas. Sin embargo, tal como ella se presenta, la consistencia es epistemológicamente muy débil y su implementación es muy problemática y lejos de ser clara. En particular, la elección de la palabra consistencia para representar este parámetro no es oportuna. En su libro sobre la evaluación de la evidencia, Robertson y Vignaux critican fuertemente su uso por los expertos forenses:

“Lo peor de todo es la palabra ‘consistencia’, una palabra desafortunada de uso común por científicos forenses, patólogos y abogados ... Contrariamente a la existencia de una comunicación clara ... los abogados generalmente interpretan ‘consistente con’ con el significado de ‘razonable fuerte apoyo’”, Robertson y Vignaux (1995: 56).

Sospechamos que los Jurados, y quizá incluso los Jueces, probablemente interpreten también *ser consistente con la procedencia de un mismo locutor* como significando *probablemente procedente de un mismo locutor*, aunque pensamos que dentro del marco de la Declaración está claro que no es lo que se entiende por el término *consistente*. Robertson y Vignaux también subrayan, como significado propio (aunque no frecuente), que consistente tiene poca fuerza epistemológica, pues ‘consistente con H’ no conlleva decir nada acerca de la probabilidad de que H sea cierta (decir que *los hechos son consistentes con H pero improbables consecuencias de H* es coherente con el significado del término *consistente*).

3.5 Problemas con poblaciones y muestras.

Como se dijo anteriormente, la Declaración de posicionamiento del Reino Unido dice que una evaluación de la evidencia del tipo subrayado en la sección 3.1 basada en la

relación de verosimilitud, aunque deseable, no es posible debido a dos factores: “problemas a la hora de definir las poblaciones de referencia relevantes”, y “falta de datos demográficos” (página 142 §6). Claramente reconocemos que son problemas reales: probablemente los más urgentes hasta el momento. El primero es teórico y se relaciona con la elección de la población relevante para la muestra, y con el tamaño de esa muestra; el segundo es práctico, y se relaciona con la posibilidad de recabar datos. Más abajo tratamos ambos problemas en orden. Argumentaremos que esos problemas, aunque son reales, no impiden el uso de las relaciones de verosimilitud.

3.5.1 La población de referencia apropiada.

Uno de los problemas mencionados por la Declaración de posicionamiento del Reino Unido sobre la utilización de enfoques basados en la relación de verosimilitud es el concerniente a la “definición de las poblaciones de referencia relevantes”. Para estimar la fuerza de la evidencia con una relación de verosimilitud, se necesita una muestra de *referencia* o de *contexto* de la población relevante⁴.

La Declaración de posicionamiento del Reino Unido acierta al asumir que se trata de un problema, y no sólo para el habla: continúa siendo un reto en la evaluación de las muestras de ADN (Aitken, 1991). La elección de la población apropiada a la muestra es, estrictamente hablando, dependiente de la hipótesis alternativa. En un caso típico, el Fiscal mantendrá la hipótesis de que la voz en la muestra cuestionada es la misma que la de la muestra conocida, es decir, que ambas son voces del imputado. La defensa mantendrá la hipótesis de que la voz en la muestra cuestionada no es la del imputado. Sin embargo, la disensión no es probable que sea simplemente que la voz cuestionada provenga de otra persona. Por el contrario, lo que se mantiene implícitamente con mayor probabilidad es que la voz cuestionada es de otro locutor, del mismo sexo que el imputado, que habla el mismo dialecto dentro del mismo idioma. La hipótesis de la defensa pudiera ser, incluso, más restrictiva: que la voz cuestionada provenga de alguien que (para un oyente no experto como un oficial de policía o un abogado) suene como la del imputado; o pudiera ser que la fuente de la voz cuestionada fuera el hermano del imputado; o incluso de un gemelo univitelino (véase la discusión en Lucy 2005: 129-133, con respecto a los límites de la población). Es probable que, en el momento de realizar el análisis, el experto en comparación de voz forense no conozca las especificidades de la hipótesis de la defensa, y se vea obligado a anticiparla. Las decisiones sobre la población de referencia relevante tendrán que tomarse caso a caso (Evetts y Weir, 1998: 44-45; ver también una discusión en Aitken 1991).

Una pregunta frecuente sobre la muestra de la población de referencia, y que puede haberse considerado implícitamente en las observaciones de la Declaración de posicionamiento del Reino Unido sobre las poblaciones es la siguiente: ¿ qué tamaño debe tener la muestra poblacional ? . Se trata de una cuestión muy importante, tanto desde el punto de vista de la investigación científica como desde el punto de vista del Tribunal o Jurado (puesto que se trata de una de las cosas que debe ser justificada). La respuesta es que depende de la precisión que tengamos que alcanzar. Aitken (1991) afronta trata sobre algunas formas de resolver este problema.

⁴ Desafortunadamente, los términos de población de referencia o población de contexto se utilizan algunas veces en la literatura científica cuando realmente son una muestra poblacional. Una población de referencia o de contexto no es lo mismo que la población de todos los posibles autores de un hecho (esta última puede tenerse en cuenta en la estimación de las apuestas a priori).

3.5.2 Falta de información sobre la distribución.

Por falta de datos demográficos asumimos, sobre la base del siguiente entrecomillado, que se quiere decir que falta conocimiento de la distribución de las propiedades de comparación forense típicas en la población.

“de la inmensa mayoría de las propiedades de voz y habla examinadas en el caso, se desconoce, sencillamente, en qué medida están distribuidas en la población” (página 140, §3).

La Declaración de posicionamiento del Reino Unido se hace eco de Rose (2002: 78) al subrayar las consecuencias de esto – que si no se conoce la incidencia de una propiedad en la población, no es posible calcular la probabilidad de observarla aleatoriamente en esa población, de este modo: “los fonetistas forenses no pueden realizar afirmaciones numéricas de probabilidad” (página 140, §3).

Resulta claramente cierto que si no tenemos acceso a estimaciones de las distribuciones de las propiedades del habla en la población relevante, no podemos estimar cuantitativamente una relación de verosimilitud (con los apropiados intervalos de confianza/límites de credibilidad) en una comparación de voz forense. Sin embargo, teniendo en cuenta lo que acabamos de afirmar, debe también seguirse que sin tales estimaciones no se pueden tomar decisiones con respecto al factor de *peculiaridad* del marco de la Declaración. No está claro cómo ha de resolverse esta inconsistencia.

Quizá, los autores y firmantes de la Declaración de posicionamiento del Reino Unido hayan considerado que los juicios sobre *peculiaridad* pudieran ser hechos por un experto basándose en sus propias experiencias, pero entonces, esa inferencia sería capaz también de estimar una relación de verosimilitud: un valor de una propiedad juzgado en la opinión de un experto como “excepcionalmente peculiar” podría también juzgarse como de baja probabilidad en la población (el denominador de la relación de verosimilitud).

La Sección 6 de la Declaración de posicionamiento del Reino Unido concluye con lo siguiente:

“Sin embargo, consideramos que la falta de datos demográficos junto con los problemas de definición de las poblaciones de referencia relevantes son los factores que impiden la aplicación cuantitativa de este tipo de enfoque en el presente contexto. En vista de estas dificultades, el marco que respaldamos es el que exponemos en 4, más arriba” (página 142, §6).

Asumimos que “el presente contexto” debe leerse como “presente contexto en el Reino Unido”, aunque esto no resulta enteramente claro. La inmensamente compleja situación lingüística actual en el Reino Unido es tal que, dada la identificación de un acento en las muestras del criminal y del sospechoso, no existe, hasta el momento, ninguna o casi ninguna información disponible sobre la distribución de las propiedades (acústicas) en ese acento. Bajo estas circunstancias, no son posibles ni una relación de verosimilitud cuantitativa ni estimaciones sobre *peculiaridad*, a menos que los expertos forenses vayan y recojan muestras de la población. Sin embargo, en Australia y en España, los

científicos forenses han recogido datos para las distribuciones y han presentado evidencias de comparación forense de voz en forma de relación de verosimilitud en los informes periciales ante los Tribunales. Reconocemos que esto está facilitado por el hecho de que la variación de acentos de inglés australiano es mucho menor que la de los acentos ingleses en el Reino Unido.

En el contexto del Reino Unido el problema permanece mientras no existan las bases de datos apropiadas de grabaciones de habla, por lo que el experto forense tendrá que conseguirlas y analizarlas. Con un sistema completamente automático, el análisis es una operación relativamente barata, pero los enfoques tradicionales fonético-acústicos requieren muchísimo más tiempo de trabajo humano para recoger y analizar los datos de referencia. Queremos hacer notar que ya se está trabajando en el Reino Unido para conseguir bases de datos y medir en ellas las propiedades (Nolan y otros, 2006; Hudson y otros, 2007; Clermont y otros, 2008); sin embargo, parece que a corto plazo en el Reino Unido no será posible realizar estimaciones cuantitativas de la fuerza de la evidencia en algunos - quizá en muchos - casos.

Una solución a corto plazo pudiera ser la de presentar estimaciones cualitativas de la fuerza de la evidencia en lugar de cuantitativas, pues sería mejor realizar tales afirmaciones en la forma de relación de verosimilitud que mediante el marco de la Declaración del Reino Unido. Tal solución ha sido recomendada por Robertson y Vignaux, y por Jessen, siempre que no estén disponibles los datos de distribución poblacionales:

“Para calcular una relación de verosimilitud no es esencial disponer de números precisos para cada una de las probabilidades. El valor de la evidencia depende de la relación entre esos números. Por consiguiente, si creemos que la evidencia es 10 veces más probable bajo una hipótesis que la otra, la relación de verosimilitud será 10, cualesquiera que sean los valores precisos del numerador y del denominador. Frecuentemente seremos capaces de valorar esta relación aproximadamente sobre la base de nuestros conocimientos y experiencias” (Robertson y Vignaux (1995: 21).

“Incluso en áreas donde no exista tal población estadística, y por tanto no sea posible una cuantificación, el enfoque bayesiano debe utilizarse como un marco conceptual que proporciona la columna vertebral del análisis de comparación de voces” (Jessen 2008: 13).

Claramente, muchos de los expertos en voz del Reino Unido son fonetistas altamente competentes con gran experiencia en comparación de voces, tanto en un contexto forense como en otros posibles, con respecto a muchas propiedades. Se puede esperar de ellos que tengan una buena intuición sobre las magnitudes esperadas de las diferencias intralocutor e interlocutor en esas propiedades. De este modo, podría esperarse que fueran capaces de proferir una opinión de experto del siguiente tenor:

“Desde mi experiencia, pienso que sería mucho más probable observar las diferencias que he pormenorizado entre las muestras de habla del criminal y del sospechoso asumiendo la procedencia de un mismo locutor que de locutores diferentes” (o *mutatis mutandis*). Una afirmación cualitativa como esta, sobre la probabilidad de la evidencia bajo hipótesis competitivas, seguramente tendría algún valor para el Tribunal, y sería coherente con nuestros argumentos sobre el marco de la relación de verosimilitud como

el lógicamente correcto para la estimación de la fuerza de la evidencia. Estamos de acuerdo, sin embargo, con Lucy cuando dice que una afirmación cualitativa es una sustitución empobrecida de una relación de verosimilitud cuantitativa (y de sus límites de credibilidad/intervalos de confianza):

“No sería ideal, ni deseable, la utilización de tales estimaciones para evaluar una evidencia clave en un juicio de un crimen importante” (Lucy 2005: 137).

4 Resumen & Conclusiones.

Esta respuesta ha criticado las propuestas esbozadas en la Declaración de posicionamiento del Reino Unido para un cambio de enfoque en la comparación de locutores forense. Los autores y firmantes de la Declaración de posicionamiento del Reino Unido merecen crédito al reconocer la necesidad de un cambio, e iniciar uno concreto en el Reino Unido. Específicamente, vemos positiva la recomendación de la Declaración de evitar las conclusiones tradicionales con problemas de lógica formal expresadas en forma de probabilidades de hipótesis, dada la evidencia; así como que hay que tener en cuenta la similitud y la rareza de las muestras de habla que se comparen. También aprobamos el uso del término *comparación* en lugar de *identificación, verificación o reconocimiento*.

Hemos identificado tres básicos puntos débiles en la propuesta. Primeramente, la aparente consideración del habla como poseedora de propiedades discretas e invariantes, a semejanza del ADN, tiene como resultado que el enfoque sería muy difícil de implementar. En segundo lugar, subrayamos la incoherencia de, por un lado, prohibir *las probabilidades de las hipótesis, dada la evidencia* y, por otro, permitirlo; así como proponer afirmaciones sobre peculiaridad mientras se reconoce la ausencia de información sobre las que basarse. En tercer lugar, la ausencia de una forma de relacionar las afirmaciones sobre consistencia y peculiaridad no ayuda al Jurado o al Tribunal a interpretarlas.

Dado el respaldo en principio de la Declaración de posicionamiento del Reino Unido del enfoque de la relación de verosimilitud, hemos intentado destacar sus aspectos positivos y evitar desacreditarlo por no ser un enfoque de relación de verosimilitud. Sin embargo, la valoración bipartita no es una relación de verosimilitud y es la relación de verosimilitud la que caracteriza el moderno pensamiento sobre la evaluación de la evidencia de comparación forense. Por consiguiente, rechazaríamos lo que se defiende en el prefacio de la Declaración de que el marco de la Declaración es “en el nivel conceptual, idéntico al que se utiliza hoy día en la presentación de la evidencia de ADN” (página 138). Consecuentemente, argumentaríamos que la Declaración de posicionamiento del Reino Unido no ha alcanzado su objetivo, sin embargo laudable, de “... llevar la disciplina de [comparación forense de voz] dentro del pensamiento moderno en otras áreas de la ciencia forense” (página 137).

Será interesante ver la propuesta del marco de la Declaración implementada en investigación y resolución de casos. Nosotros, desde luego, animaríamos a los investigadores y expertos en comparación de voz forense a que rápidamente se adaptaran a realizar afirmaciones de relación de verosimilitud cuantitativas como estándar (aunque habrá algunas propiedades de las muestras de habla forenses sobre las que sólo será posible dar estimaciones cualitativas). Dada la cantidad de energía que en

el Reino Unido se está desplegando para conseguir datos cuantitativos para sustentar la comparación de voz forense basada en relaciones de verosimilitud, esperamos que eso sea pronto una realidad en el Reino Unido.

Referencias

- Aitken, C. G. G. and Stoney, D. A. (1991) *The Use of Statistics in Forensic Science*. Chichester, UK: Ellis Horwood.
- Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Forensic Evidence for Forensic Scientist* (2nd ed.). Chichester, UK: Wiley.
- Aitken, C.G.G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Applied Statistics*, 53(4):109-122.
- Balding, D. J. (2005) *Weight-of-evidence for Forensic DNA Profiles*. Chichester, UK: Wiley.
- Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53: 370–418.
- Champod, C. and Meuwly, D. (2000) The inference of identity in forensic speaker recognition. *Speech Communication*, 31: 193–203.
- Clermont, F., French, J. P., Harrison, P., and Simpson, S. (2008) Population data for English Spoken in England. Paper presented at the 17th meeting of the International Association for Forensic Phonetics and Acoustics, Lausanne, Switzerland, July 2008.
- Donnelly, P. (2005) Appealing statistics. *Significance*, 2(1): 46–48.
- Evetts, I. W. (1977) The interpretation of refractive index measures. *Forensic Science International*, 9: 209–217.
- Evetts, I. W. (1991) Interpretation: A personal odyssey. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 9–22. Chichester, UK: Ellis Horwood.
- Evetts, I. W. (1998) Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3): 198–202.
- Evetts, I. W., and Weir, B. S. (1998) *Interpreting DNA Evidence*. Sunderland, MA: Sinauer Associates.

- Foreman, L. A., Champod, C., Evett, I. W., Lambert, J. A., and Pope, S. (2003) Interpreting DNA evidence: A review. *International Statistics Journal*, 71: 473–473
- French, J. P. & Harrison, P. (2007) Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law*, 14(1): 137–144
- Friedman, R.D. (1996) Assessing evidence. *Michigan Law Review*, 94: 1810–1838.
- González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., and Ortega-García, J. (2006) Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2-3): 331–355.
- González-Rodríguez, J., Rose, P., Ramos, D., Torre, D., and Ortega-García, J. (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7): 2104–2115.
- Good, I. J. (1991) Weight of evidence and the Bayesian likelihood ratio. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 85–106. Chichester, UK: Ellis Horwood.
- Haigh, J. (2005) Review of statistics and the evaluation of evidence for forensic scientists. *Significance*, March: 40.
- Hodgson, D. (2002) A Lawyer looks at Bayes' Theorem. *The Australian Law Journal*, 76: 109–118.
- Hudson, T., de Jong, G., McDougall, K., Harrison, P., Nolan, F. (2007) F0 statistics for 100 young male speakers of standard Southern British English. In J. Trouvain and W. J. Barry (eds.) *Proceedings of the 16th International Congress of Phonetic Sciences* 1809–1811.
- Jessen, M. (2008) Forensic phonetics. *Language and Linguistics Compass*, 2(4): 671–711.
- Lindley, D. V. (1991) Probability. In C. G. G. Aitken and D. A. Stoney (eds.) *The Use of Statistics in Forensic Science* 27–50. Chichester, UK: Ellis Horwood.

- Lucy, D. (2005) *Introduction to Statistics for Forensic Scientists*. Chichester, UK: John Wiley.
- Nolan, F. (1983). *The phonetic Bases of Speaker Recognition*. Cambridge, UK: Cambridge University Press.
- Nolan, F. (1996) Forensic phonetics. Notes distributed at the two-week course at the 1996 Australian Linguistics Institute, Australian National University, Canberra.
- Nolan, F. (1997) Speaker recognition and forensic phonetics. In W. J. Hardcastle and J. Laver (eds) *The Handbook of Phonetic Sciences* 744–767. Oxford, UK: Blackwell.
- Nolan F., McDougall, K de Jong, G., and Hudson, T. (2006) A Forensic Phonetic Study of 'Dynamic' Sources of Variability in Speech: The DyViS Project. In Warren & Watson (eds.) *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* 13–18.
- Pigeon, S., Druyts, P., and Verlinde, P. (2000) Applying logistic regression to the fusion of the NIST '99 1-speaker submissions. *Digital Signal Processing*, 10: 237–248.
- Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10: 19–24.
- Robertson, B. and Vignaux, G.A. (1995) *Interpreting Evidence*. Chichester, UK: Wiley.
- Rose P (2002) *Forensic Speaker Identification*. London & New York: Taylor and Francis.
- Rose P (2003) *The Technical Comparison of Forensic Voice Samples*. Issue 99, *Expert Evidence*. Freckelton I, Selby H, (series eds.). Sydney, Australia: Thomson Lawbook Company.
- Rose (2005) Forensic Speaker Recognition at the beginning of the Twenty-First century. An overview and a demonstration. *Australian Journal of Forensic Sciences*, 37(2): 49–71.
- Saks, M. J., and Koehler, J. J. (2005) The coming paradigm shift in forensic identification science. *Science*, 309: 892–895.
- Thompson, W. C., and Schumann, E. L. (1987) Interpretation of statistical evidence in criminal trials. The prosecutor's fallacy and the defence attorney's fallacy. *Law and Human Behaviour*, 11: 167-187.